

Numerical analysis of parabolic optimal control problems with restrictions on the state variable and its first derivative

vom Fachbereich Mathematik der Technischen Universität Darmstadt
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
- Dr. rer. nat. -
genehmigte Dissertation

Tag der Einreichung: 31.05.2017
Tag der mündliche Prüfung: 03.08.2017

Referent: Prof. Dr. Winnifried Wollner
Korreferent: Prof. Dr. Boris Vexler

von

Dipl.-Math. Francesco Ludovici

aus L'Aquila, Italien

Darmstadt, D 17
2017

Abstract

The aim of this thesis is the numerical analysis of optimal control problems governed by parabolic PDEs and subject to constraints on the state variable and its first derivative. The control is acting distributed in time only while the state constraints are considered point-wise in time and global in space; this setting generates an optimization problem of semi-infinite type.

The consideration of a space-time discretization of the problem requires the analysis of the convergence of the discretized solution toward the continuous one, as temporal and space mesh size tend to zero. This is based, at any level of discretization, on a priori error estimates for the solution of the parabolic differential equation which are obtained within this thesis.

One of the main challenge for state-constrained problem consists in the presence of a Lagrange multiplier appearing as a Borel measure in the system of first-order optimality conditions. In particular, such measure enters the optimality system as data in the adjoint equation affecting the regularity of the adjoint variable itself. Therefore, in the derivation of the convergence rates the use of adjoint information has to be avoided. When considering non-convex problems, the presence of local solutions and the need for second-order optimality conditions require a different strategy compared to the convex case, making the analysis more involved. In particular, the convergence of the discretized solution toward the continuous one is based on a so-called quadratic growth-condition, which arises from the second-order optimality conditions. The a priori error estimates for the PDEs are verified numerically.

Zusammenfassung

Ziel der vorliegenden Dissertation ist die numerische Analysis von Optimalsteuerungsproblemen mit parabolischen Differentialgleichungen sowie Zustands- und Gradientenschranken als Nebenbedingungen. Während die Steuerung lediglich zeitabhängig ist, werden die Zustandsschranken punktweise in der Zeit und global im Ort vorgeschrieben. Es handelt sich somit um ein semi-infinites Optimierungsproblem.

Die Diskretisierung des Problems in Ort und Zeit erfordert eine Konvergenzanalyse, d.h. eine Betrachtung des Verhaltens der Lösungen der diskretisierten Probleme bezüglich der Lösung des kontinuierlichen Problems, wenn die Gitterweite der Diskretisierung gegen Null strebt. Diese Konvergenzanalyse basiert bei beliebigem Diskretisierungsgrad auf a-priori-Fehlerschätzern für die parabolischen Differentialgleichungen, welche in der vorliegenden Dissertation hergeleitet werden. Eine der großen Herausforderungen bei Problemen mit Zustandsschranken besteht darin, dass es sich bei den Lagrange-Multiplikatoren in den Optimalitätsbedingungen erster Ordnung um Borelmaße handelt. Insbesondere treten diese Maße als Daten in der Adjungiertengleichung auf, wodurch sie die Regularität des adjungierten Zustands direkt beeinflussen. Folglich muss in der Herleitung der Konvergenzraten auf Informationen aus der Adjungiertengleichung verzichtet werden. Das Auftreten lokaler Optima und die Verwendung von Optimalitätsbedingungen zweiter Ordnung bei der Untersuchung nicht-konvexer Probleme erfordert die Wahl einer anderen Strategie als im konvexen Fall, sodass die Komplexität des vorliegenden Problems erheblich zunimmt. Insbesondere basiert die Konvergenzanalyse auf einer quadratischen Wachstumsbedingung, welche aus den Optimalitätsbedingungen zweiter Ordnung hervorgeht.

Die A-priori-Fehlerschätzer für die partiellen Differentialgleichungen werden anhand numerischer Beispiele verifiziert, wobei die Konvergenzraten des Optimalsteuerungsproblems den entsprechenden Fehler in der Differentialgleichung erwartungsgemäß widerspiegeln.

Acknowledgments

First and foremost I would like to offer my sincere thanks to my thesis advisor Prof. Dr. Winnifried Wollner. His mentorship, support and guidance along the last years have been invaluable and source of inspiration.

This thesis would not have been possible without the support of my family, Carla, Lino, Alessandro, Diana, and the complicity many years ago of Ada, Gina, Antonio, Andrea.

In addition, my gratitude goes to Prof. Dr. Ira Neitzel for a fruitful collaboration, to Prof. Dr. Boris Vexler for some constructive discussions during his visit in Hamburg, and to Prof. Dr. Eduardo Casas for a brief though profitable discussion.

Further, I would like to thank all the colleagues and academic staff I worked and collaborated with at Universität Hamburg and TU Darmstadt. In particular, my sincere thanks to Johann Schmitt for his help with the German part of this thesis.

To my friend Leo and his cat, and to Antonio, Dani, Mattia, Pedro.

Last but not least, I thank my high-school math teacher who strongly suggested me to not study Mathematics.

Contents

1	Introduction	3
2	Fundamentals and problem formulation	9
2.1	Basics and notation	9
2.2	Problem setting	10
2.2.1	State Equation	10
2.2.2	Optimization Problem	11
2.2.3	Discretization	14
2.3	Examples	17
2.3.1	First-order state constraint with linear state equation . .	17
2.3.2	Zero-order state constraint with semi-linear state equation	18
3	Parabolic partial differential equations	21
3.1	Well-posedness and regularity of solutions for linear equations . .	22
3.2	Semi-linear differential equations	24
3.3	Differential equations with measure as data	27
3.4	Differentiability of the functionals	28
4	A priori error estimates for the state equation	33
4.1	Auxiliary problems and the error analysis	35
4.2	Stability analysis	40
4.2.1	Linear problem	40
4.2.2	Extension to semi-linear problems	42
4.3	Error analysis for linear equations	49
4.3.1	Temporal error	49
4.3.2	Spatial error	59
4.4	Error Analysis for semi-linear equations	65
4.4.1	Temporal error	65
4.4.2	Spatial error	75
4.4.3	Extension to first-order integral constraints	84
4.5	Numerical results	85
5	Optimization problems with state constraints	89
5.1	Optimality conditions	91
5.1.1	First order optimality conditions	92
5.1.2	Second order optimality conditions	96
5.2	Convergence analysis for convex optimization problems	103
5.2.1	First-order integral constraints pointwise in time	103
5.3	Convergence analysis for non-convex optimization problems . . .	108
5.3.1	Zero-order integral constraints pointwise in time	108
6	Conclusions and outlook	117

1. Introduction

The focus of this thesis is the numerical analysis of a class of optimization problems governed by a parabolic partial differential equation (PDE). The problems are subject to constraints on the control variable and, most importantly, restrictions on the state variable and its first derivative.

The underlying differential equation can include a semi-linear term, giving rise to a non-convex optimization problem.

A space-time discretization of the problem is considered leading naturally to the central matter of this work, the derivation of convergence rates for the error between the continuous optimal solution and its discrete counterpart, in terms of the temporal and spatial mesh size.

The class of problems at hand consists in the minimization of a convex cost functional

$$\min J(q, u)$$

where the control q belongs to a closed convex set Q_{ad} reflecting the presence of bi-lateral inequality constraints.

The control acts in time only and it is represented by a \mathbb{R}^m -vector. Control and state variable are coupled by a parabolic differential equation

$$\partial_t u(t, x) + \mathcal{A}u(t, x) + d(t, x, u) = q(t)g(x)$$

where \mathcal{A} is an elliptic differential operator, g is a fixed function, and the Nemytskii operator $d(t, x, u)$ reflects the presence of a semi-linearity. Further, suitable initial and boundary conditions are prescribed.

Additionally, and most importantly, the state variable and its first derivative are subject to restrictions of the type

$$\int_{\Omega} u(t, x)\omega(x)dx \leq 0 \quad \forall t \in [0, T],$$

and

$$\int_{\Omega} |\nabla u(x, t)|^2 \omega(x)dx \leq 0, \quad \forall t \in [0, T]$$

where ω is a weighting function.

The innovative part of this work is the consideration of gradient state constraints in the context of convex optimization problems, and, more generally, the consideration of state constraints for the non-convex setting. Indeed, while more attention has been recently given to the discussion of optimality conditions for this class of problems, the numerical analysis is still at an early stage, particularly for non-convex problem.

State constraints point-wise in time and global in space, together with the control being a \mathbb{R}^m -vector acting in time, generate an optimization problem of semi-infinite type. As it will be clear from the subsequent analysis, this class of problems is meaningful from a point of view of real-world applications and

challenging from a mathematical perspective.

In particular, regarding the difficulties of the problem at hand, to the before-mentioned state constraints correspond Lagrange multipliers which are Borel measures. This is the case because the state constraints are formulated in suitable spaces of continuous functions. As a consequence, the adjoint equation in the system of first-order optimality conditions presents a measure as data which afflicts the regularity of the adjoint variable. This issue must be considered when deriving convergence rates, as we cannot rely on adjoint information.

This situation is magnified in the non-convex setting, where, due to the presence of local solutions, a different approach is required. While in the convex case the first-order optimality conditions are necessary and sufficient for (global) optimality, in the non-convex case second-order conditions need to be postulated in order to ensure (local) optimality.

Additionally, since the problem is discretized in time and space, it is clear that a priori error estimates must be derived in order to guarantee the convergence of the solution of the discrete problem toward the solution of the continuous one. The difficulty here lies in the fact that, to treat state constraints point-wise in time, estimates in the L^∞ -norm with respect the temporal variable are necessary. In particular, the zero-order and first-order state constraints require estimates in the $L^\infty(I, L^2(\Omega))$ and $L^\infty(I, H_0^1(\Omega))$ -norm, respectively.

We now unveil the structure of this thesis. To tackle the problems discussed above, we prepare the ground in the first two chapters introducing the tools which will be used in the rest of this work. In Chapter 2, the notation is fixed and the main parts of an optimal control problem are described in a formal and abstract way. The chapter is concluded with Section 2.3 where the problems under consideration are stated concretely.

In Chapter 3, we recollect results on the regularity of the solutions of linear and semi-linear parabolic PDEs from classical monographs. It is fundamental that the regularity of the data of the differential equation permits the embedding of the resulting state space into a space of continuous functions. This is needed to guarantee the state constraint to be well-posed. From the structure of the state constraints, point-wise in time and averaged in space, this means that the sought spaces are $C(\bar{I}, L^2(\Omega))$ and $C(\bar{I}, H_0^1(\Omega))$ for the zero-order and first-order state constraints, respectively. A question strictly related to the last point is the regularity of the solution when a measure enters the problem as data. Indeed, this happens when formulating the first-order optimality conditions, as the Lagrange multiplier associated with a state constraint point-wise in time lies in the dual space of $C(\bar{I})$. This question will be addressed in Section 3.3 where we give an insight into the subject. Further, in the last section of the chapter, we introduce several functionals and operators whose differentiability properties depend upon the regularity of the solution of the PDEs.

The core of this thesis consists of Chapter 4. It is here that we obtain the a priori error estimates for the PDEs which will constitute the base for the derivation of the convergence rates for the optimal control problems. The influence of the temporal and spatial discretization is considered separately. The derivation is based on a duality argument requiring, at any level of discretization, error estimates in negative norms for some associated dual problems; the error under consideration in the primal variable represents the initial condition of the auxiliary problem. Therefore, by prescribing additional regularity to the

initial condition of the primal problem, we can fully exploit the approximation properties of the space-time discretization. These are analyzed in Section 4.2, where we obtain stability estimates for the solutions of the auxiliary problems. Among others, time-weighted estimates will be required in order to handle the L^∞ -norm in time.

In a second step, we obtain in Section 4.3 and 4.4 the negative norm estimates for the auxiliary problems. This will be done through a splitting of the total error into a projection and a discretization error. For the former, and in the time-discrete setting, we exploit the error arising when truncating the time interval in the dual auxiliary problem. Then, the time-weighted estimate, and the approximation properties of the projection operator employed, will provide a bound for the projection error. For the spatial discretization, such bound will be provided by the usual L^2 -projection. The discretization error will be investigated using the before-derived approximation properties of the space-time approximation, plus additional estimates for some forward auxiliary problems. The latter are required to deal with the error at the final time of the dual problem.

Though similar in the structure as the linear setting, the treatment of the semi-linear equation will require more effort. To deal with the semi-linear term, a linearization is performed by introducing an $L^\infty(I \times \Omega)$ term; this will lead to additional auxiliary problems with related estimates, namely in the $L^2(I, L^2(\Omega))$ -norm. The chapter is concluded with the numerical validation of our findings.

In Chapter 5, we assemble the results derived in the previous chapters to obtain the rate of convergence of the discretized solution of the optimal control problem. Firstly, we review the optimality conditions for the two problems at hand. The first-order conditions are necessary and sufficient for optimality in the convex case, while the treatment of the non-convex problem requires the formulation of second-order sufficient conditions (SSCs). Rather than relying on a stronger SSC, we use a weaker SSC based on a cone of critical directions inherited from the finite-dimensional optimization theory. From this discussion, it is clear that to derive rates of convergence one should use two different approaches, depending whether the problem is convex or not. Further, as already mentioned, the strategy used has to avoid the use of the adjoint variable. In the convex setting, this is achieved, in Section 5.2, exploiting the variational inequality and the complementary slackness condition from the system of first-order optimality conditions. On the other hand, the non-convex case, which is addressed in Section 5.3, requires more effort and the introduction of further localized problems. These are needed to deal with the presence of local solutions, and are used in combination with the so-called two-way feasibility in order to overcome the rough regularity of the adjoint variable. Firstly, based on the (linearized) Slater point, we construct sequences of controls which are feasible for the continuous and discrete problem. Then, these sequences are used together with the quadratic-growth condition, coming from the SSCs, to obtain the rate of convergence of the discretized solution toward the continuous one. While in the convex setting we obtain a clear separation of the temporal and spatial error, this is not possible, in general, for the non-convex case when using weaker SSCs. Indeed, this would require the transfer of the SSCs to the semi-discrete problem and, thus, the investigation of the convergence of critical directions. Therefore, the derivation is performed in one step, from the continuous to the discrete problem directly.

We now move our attention to the literature related to the topic of this thesis starting with the possible real-world applications. The class of problems under consideration represents a simplified model for industrial processes like cooling/heating in steel and glass manufacturing. Due to the high-temperature involved in these processes, the underlying differential equation is usually formulated through a non-linear parabolic equation coupled with a transport equation for the radiation intensity, [19, 72, 73]. Particularly in the glass industry, it is important to keep track of the thermal stress to avoid material failure and preserve some desired properties in the final product. Since the thermal stress is represented by the gradient of the temperature, this naturally lead to the consideration of constraints on the first derivative of the state variable. Additionally, especially in the steel industry, the cooling of the material profile is performed by finitely many water-spray nozzles fixed in space and activated in time, [26, 86, 87]. This is indeed the form of the control variable considered in this thesis. Further, this same structure of the control variable, together with restrictions on the state variable, appear in others applications spanning from crystals growth, [65], to local hypothermia in cancer treatment, [23]. For further possible applications, we refer to [43, Chapter 4], see also [20].

The literature on a priori error estimates for parabolic optimal control problems in presence of state constraints has only few contributions. Based on techniques from [57, 74], in [61] error estimates in $L^\infty(I, L^2(\Omega))$ are obtained for a linear differential equation in presence of constraints on mean values of the state. These have been extended by the author of this thesis in [56], where $L^\infty(I, H_0^1(\Omega))$ -norm estimates are derived for integral constraints on the first-derivative of the state variable, and in [55], where a semi-linear PDE is considered in a setting similar to [61]. To the best of the author's knowledge, the results in [56] and [55] are the first to perform an a priori analysis in presence of gradient constraints and state constraints with semi-linear equations, respectively.

In [22], state constraints point-wise in time and space are considered in relation with a linear PDE with time-dependent control, and a discretization scheme where time step and spatial grid are coupled. This has been extended in [35] to include control box-constraints. We mention also [88] where a state constraint at final time is prescribed.

Regarding related literature without state constraints, an important contribution to the subject has been given by [62, 63], where the authors have considered the properties of the space-time discretization employed in this thesis. In [68], the case of control box-constraints in presence of a semi-linear PDE is considered; further the authors consider several control discretization approaches. The case of control constraints has been investigated also in [36, 37, 49, 50, 78]. In [18], plain convergence is shown in presence of a semi-linear PDE.

Less sparse is the literature for elliptic PDEs. For gradient state constraints we refer to [21, 71, 90] and the reference therein. For the case of a semi-linear elliptic equation and state constraints we refer to [11, 42, 66].

More attention has received recently the study of optimality conditions for parabolic optimal control problems in presence of constraints on the state and its first derivative. Confining ourselves to the case of gradient constraints, the existence and optimality conditions for an integral constraint point-wise in time has been addressed in [59]. Using Ekeland's variational principle, in [12]

the authors have obtained a Pontryagin's principle for an integral gradient constraint, while in [76] second order sufficient conditions are investigated for several constraints including integral gradient restrictions. A thorough analysis of SSCs in presence of integral gradient constraints has been investigated in [60]. Concerning, more generally, integral constraints on the state variable, seminal papers for the theory of SSCs are [7, 33]. In the former, the Ekeland's principle is used in a setting with boundary controls and different types of state constraints are treated. The latter deals with a nonlinearity in the boundary condition and uses the semi-group theory to overcome some limitations in the dimension of the problem. The results of both papers have been refined in [8], where the authors borrowed techniques from nonlinear optimization in finite-dimensional spaces to obtain SSCs close to the necessary one. This approach, limited in a first place to the case $\Omega \subset \mathbb{R}^1$, has been extending in [20] to domains of arbitrary dimensions using time-dependent controls represented by an \mathbb{R}^m -vector. Indeed, this is the approach employed in this thesis. Time-dependent controls and integral state constraints point-wise in time have been considered also in [5] when a cubic non-linear term enters the state equation.

Other relevant contributions to the theory of SSCs in presence of state constraints are [3, 13, 15, 16, 47, 75, 77]. Lastly, we also mention the survey [17] on SSCs for parabolic optimal control problem where the interested reader can find also the main references for the elliptic case.

Further specific references will be given throughout the thesis.

2. Fundamentals and problem formulation

In this chapter, the notation is fixed and the main features of the problem under consideration are depicted in abstract form. All the details and the specific setting of our problem, will be given in the others chapters of this thesis. Section 2.1 is focused on the notation used more often throughout this thesis and, more generally, facts regarding Sobolev spaces with negative index and spaces depending on time.

In Section 2.2 the principal components of a parabolic optimal control problem are introduced in a concise though formal way. After introducing the state equation, we provide an overview of the optimality conditions and the discretization of the problem.

Finally, in Section 2.3, we specify concretely the abstract description of the problem with the state equations and the state constraints analyzed in this work. As it is not the aim of this thesis to recollect well-established notions of functional analysis and optimal control theory, we omit the details referring to the classic monographs [51, 53, 85] and to the references given through this section.

2.1 Basics and notation

Let $\Omega \subset \mathbb{R}^d$, with $d \in \{2, 3\}$, be a convex bounded domain with C^3 - boundary $\partial\Omega$. The problem is formulated in the time interval $I = (0, T)$. Lebesgue and Sobolev space are denoted with $L^p(\Omega)$, $W^{s,p}(\Omega)$, respectively, and for $p = 2$ we shorten $H^s(\Omega) := W^{s,2}(\Omega)$. The sub-index $W_0^{s,p}(\Omega)$ indicates that the function is zero on the boundary. The scalar product in $L^2(\Omega)$ is denoted by (\cdot, \cdot) with corresponding norm $\|\cdot\|$. For the dual of a Sobolev space we use the convention $H^{-s} := (H_0^s)^*$, $s \geq 0$ and the negative Sobolev norm

$$\|v\|_{-s} = \sup_{\varphi \in H_0^s(\Omega)} \frac{(v, \varphi)}{\|\varphi\|_{H_0^s(\Omega)}},$$

where the identification of the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ with the scalar product (\cdot, \cdot) is guaranteed by

$$(2.1) \quad H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$$

being a Gelfand triplet. When referring to the time interval I , we use $(\cdot, \cdot)_{L^2(I)}$ and $\|\cdot\|_{L^2(I)}$ for the scalar product and norm in $L^2(I)$, respectively. The space of continuous function on \bar{I} is $C(\bar{I})$ and its dual is identified with the space of Borel measure $\mathcal{M}(I)$ with duality pairing $\langle \cdot, \cdot \rangle_{C(I), C(\bar{I})^*}$.

In order to treat parabolic problems, we need spaces involving time, or more generally, vector-valued functions. For a generic Hilbert space H , we denote with $C(\bar{I}, H)$ the space of continuous function from I and values in H , endowed with norm $\|v\|_{C(\bar{I}, H)} := \max_{t \in \bar{I}} \|v(t)\|_H$. Similarly, $L^p(I, H)$, $1 \leq p < \infty$, is the set of measurable functions from I with values in H such that

$$\|v\|_{L^p(I, H)} := \left(\int_0^T \|v(t)\|_H^p dt \right)^{1/p} < \infty.$$

2.2. Problem setting

For $p = 2$, we have an Hilbert space with scalar product $(\cdot, \cdot)_I := \int_I (\cdot, \cdot)_H dt$ and norm $\|\cdot\|_I$.

In the same fashion, for $L^\infty(I, W)$ we have

$$\|v\|_{L^\infty(I, W)} := \operatorname{ess\,sup}_{t \in I} \|v(t)\|_W \leq \infty.$$

An Hilbert space which will be used often in the following is

$$W(0, T) := \{u \in L^2(I, H_0^1(\Omega)), \partial_t u \in L^2(I, H^{-1}(\Omega))\}.$$

The space of polynomials of maximum degree m on the time interval I with values in H is denoted with $\mathcal{P}_m(I, H)$.

We conclude this section with further concepts regarding negative Sobolev norms referring to [84]. To this end, we introduce the Sobolev space

$$\dot{H}^s(\Omega) = \{v \in H^s(\Omega) \mid \Delta^j v = 0 \text{ on } \partial\Omega \text{ for } j \in \mathbb{N}_0\}.$$

and the iterated solution operators for Poisson's problem

$$\begin{aligned} -\Delta^{-1} &: H^{-1}(\Omega) \rightarrow \dot{H}^1(\Omega), \\ -\Delta^{-1} &: L^2(\Omega) \rightarrow \dot{H}^2(\Omega), \\ -\Delta^{-2} &: H^{-1}(\Omega) \rightarrow \dot{H}^3(\Omega), \end{aligned}$$

observing that, thanks to the C^3 -regularity of the boundary, they are continuous operators; see, e.g., [32, Theorem 8.13].

Considering the semi-norm

$$|\cdot|_{-s} := (-\Delta^{-s} \cdot, \cdot)^{1/2},$$

it follows that this is equivalent to the negative norm of $\dot{H}^s(\Omega)$, see [84, Lemma 5.1], and therefore we can define the following equivalent norms on $H^{-s}(\Omega)$ and $L^2(I, H^{-s}(\Omega))$

$$\begin{aligned} \|\cdot\|_{H^{-1}(\Omega)} &:= \|\nabla \Delta^{-1} \cdot\|, & \|\cdot\|_{L^2(I, H^{-1}(\Omega))} &:= \|\nabla \Delta^{-1} \cdot\|_I, \\ \|\cdot\|_{H^{-2}(\Omega)} &:= \|\Delta^{-1} \cdot\|, & \|\cdot\|_{L^2(I, H^{-2}(\Omega))} &:= \|\Delta^{-1} \cdot\|_I, \\ \|\cdot\|_{H^{-3}(\Omega)} &:= \|\nabla \Delta^{-2} \cdot\|, & \|\cdot\|_{L^2(I, H^{-3}(\Omega))} &:= \|\nabla \Delta^{-2} \cdot\|_I. \end{aligned}$$

Further, we observe that from the definition of $\dot{H}^s(\Omega)$, the norm $\|\cdot\|_{H^{-s}}$ corresponds to the norm of $(H^s(\Omega) \cap H_0^1(\Omega))^*$ when $s = 1, 2$. While for $s = 3$, we have the additional condition $\Delta v = 0$; see [56, Remark 2.7].

Throughout the thesis, we denote with c and C two generic constants independent of the discretization parameters that might take different values at each appearance.

2.2 Problem setting

2.2.1 State Equation

The first component to analyze in an optimal control problem is the differential equation coupling the state variable u with the control variable q , in terms of

existence, uniqueness, and regularity of its solution. As we will see, the regularity of the solution takes on great importance in all aspects of the problem, from the discretization to the numerical analysis.

The set of admissible controls is denoted by Q_{ad} and we assume this set to be non-empty, closed, convex, and bounded. The state space is denoted by U . The former will be defined in Section 2.3 while the latter in Chapter 3. Denoting by \mathcal{A} the standard uniform symmetric elliptic operator and with ∂_t the derivative with respect the time variable, we consider the following parabolic problem

$$(2.2) \quad \begin{aligned} (\partial_t + \mathcal{A})u(t, x) &= f(t, x) = q(t)g(x) && \text{in } I \times \Omega, \\ u(t, x) &= 0 && \text{on } I \times \partial\Omega, \\ u(0, x) &= u_0 && \text{in } \{0\} \times \Omega, \end{aligned}$$

where u_0 is the initial condition and g is a given function. The setting above also can include a non-linear term expressed via a Nemytskii operator $d(t, x, u(t, x))$ which will be defined in Section 2.3.2.

The splitting of the right-hand side into a temporal and spatial part is motivated by real-world applications of the problem mentioned in Chapter 1. Additionally, this splitting will be helpful from a mathematical perspective to overcome some issues related to the dimension of the domain.

Denoting with $b(\cdot, \cdot)$ a suitable bilinear form associated with $(\partial_t + \mathcal{A})$, the weak formulation of (2.2) reads: for $q \in Q_{\text{ad}}$ find $u \in U$ such that

$$(2.3) \quad b(u, \varphi) = (qg, \varphi)_I + (u_0, \varphi(0)), \quad \forall \varphi \in U.$$

The regularity of the weak solution u of (2.3) must allow the embedding of the state space U into a space of continuous functions to ensure a regularity condition for the state constraint; we will see in Chapter 3 that this requirements is met.

After showing, in Sections 3.1 and 3.2, the well-posedness of the state equation, one introduces the control-to-state map

$$(2.4) \quad S: Q_{\text{ad}} \rightarrow U, \quad q \rightarrow S(q) = u(q)$$

associating to any given control $q \in Q_{\text{ad}}$ the solution $u(q)$ of (2.3). This operator is continuous and we will exploit the continuity to ensure the existence of a solution for the optimal control problem in the forthcoming section.

Remark 2.2.1. *To be more precise, one should distinguish two cases for the control-to-state map reflecting the fact that we regard the state variable u in two different spaces: the one appearing in the objective functional and the one in the state constraint. For the scope of this chapter the definition given above is enough. We will come back to this point once the regularity of all the components of the problem is established.*

2.2.2 Optimization Problem

We formulate an optimal control problem in abstract form explaining the main steps to guarantee its well-posedness. Once the existence of a solution has been established, we briefly discuss the role of first and second-order optimality

conditions, which will be later treated in Chapter 5.

Let \mathcal{Q} and \mathcal{Z} be two Banach spaces with the former being reflexive, and such that $Q_{\text{ad}} \subset \mathcal{Q}$, where Q_{ad} has the properties stated in the previous section. Further, we consider a closed convex cone $\mathcal{K} \subset \mathcal{Z}$ and the Frechét differentiable operators

$$j: \mathcal{Q} \rightarrow \mathbb{R}^+, \quad G: \mathcal{Q} \rightarrow \mathcal{Z}.$$

We assume the operator j to be continuous, positive, and strictly convex. Then, the following abstract problem is under consideration

$$(2.5) \quad \begin{aligned} & \text{minimize } j(q) \\ & \text{subject to } q \in Q_{\text{ad}}, \quad G(q) \in \mathcal{K} \end{aligned}$$

Remark 2.2.2. *The solution operator S of the differential equation defined in the previous section is implicitly included in the setting above. Indeed, the operator j , which plays the role of the reduced cost functional, is well-posed thanks to the continuity of the control-to-state map. Further, we will see that the operator G is the concatenation of the operator defining the state constraint with the control-to-state map. Then, it is clear that when talking about convex and non-convex problems we are referring to the problem above with respect to the solution operator of the linear and non-linear differential equation, respectively.*

The first natural step is to specify the meaning of solution. Indeed, in the case of a convex problem, we clearly expect a global solution while for a non-convex problem the concept of local solutions comes into play.

Definition 2.2.3. *Let $\bar{q} \in Q_{\text{ad}}$ be a control such that $G(\bar{q}) \in \mathcal{K}$, and let X be a generic Banach space equipped with $\|\cdot\|_X$. We call \bar{q} a global optimal solution if*

$$j(\bar{q}) \leq j(q), \quad \forall q \in Q_{\text{ad}} \text{ with } G(q) \in \mathcal{K}.$$

If there exists some $\epsilon > 0$ such that the above relation is satisfied for all $q \in Q_{\text{ad}}$ with $\|q - \bar{q}\|_X \leq \epsilon$ and $G(q) \in \mathcal{K}$, then we call \bar{q} a local solution in the sense of X .

The proof of existence of a solution for (2.5) can be regarded as a standard procedure. We sketch the main steps hereafter referring for all the details to [51, Chapter 3, Section 2], see also [43, Theorem 1.43], for the convex case and [51, Theorem 15.1], see also [85, Theorem 5-7], for the non-convex setting.

Theorem 2.2.4. *Let assume that (2.5) possesses a feasible point. Then, if (2.5) is strictly convex there exists a unique global solution \bar{q} . If the problem is non-convex there exists at least a local solution \bar{q} .*

Proof. We start with the convex case. The positiveness of j and Q_{ad} being non-empty guarantee the existence of a minimizing sequence $\{q_n\} \subset Q_{\text{ad}}$. Since Q_{ad} is bounded, also $\{q_n\}$ is bounded and therefore we can extract a weakly convergent subsequences $q_{n_i} \rightharpoonup \bar{q}$ because \mathcal{Q} is reflexive. Since j is continuous and convex with domain in a Banach space, it is weakly-lower semi-continuous, which, in turn, guarantees the optimality of \bar{q} . The uniqueness easily follows due to j being strictly convex.

The proof for the non-convex case differs in a few points related to the presence of the Nemytskii operator $d(t, x, u(q))$ in the state equation. We come back to this point in Section 5.1.1 when all the necessary tools will be defined. \square

In a next step, we seek for conditions ensuring the optimality of \bar{q} . We introduce the Lagrangian functional associated with (2.5)

$$(2.6) \quad \mathcal{L}: Q_{\text{ad}} \times \mathcal{Z}^* \rightarrow \mathbb{R}, \quad \mathcal{L}(q, \mu) = j(q) + \langle G(q), \mu \rangle_{\mathcal{Z}, \mathcal{Z}^*},$$

and

$$(2.7) \quad \mathcal{K}^+ = \{ \mu \in \mathcal{Z}^* \mid \langle v, \mu \rangle_{\mathcal{Z}, \mathcal{Z}^*} \geq 0, \forall v \in \mathcal{K} \}$$

is the dual cone associated with \mathcal{K} , where μ is a Lagrange multiplier and $\langle \cdot, \cdot \rangle_{\mathcal{Z}, \mathcal{Z}^*}$ is the duality pairing between the space \mathcal{Z} and its dual. For the definition of Lagrange multiplier in this context see, e.g., [45, Definition 1.1]. We assume the Lagrangian to be twice differentiable, as it will be the case.

When the problem is convex the first-order optimality conditions are necessary and sufficient for optimality. These are expressed through a variational inequality and a complementary slackness condition, with the existence of the Lagrange multiplier ensured by the so-called Slater's regularity condition. We summarize this in the following.

Theorem 2.2.5. *Let \bar{q} be the optimal solution of (2.5) and \tilde{q} be a feasible control such that*

$$(2.8) \quad G(\tilde{q}) \in \text{int } \mathcal{K}.$$

Then there exists a Lagrange multiplier $\bar{\mu} \in \mathcal{K}^+$ associated with \bar{q} such that there holds the variational inequality

$$(2.9) \quad j'(\bar{q})(q - \bar{q}) + \langle G'(\bar{q})(q - \bar{q}), \bar{\mu} \rangle_{\mathcal{Z}, \mathcal{Z}^*} \geq 0,$$

and the complementary slackness condition

$$(2.10) \quad \langle G(\bar{q}), \bar{\mu} \rangle_{\mathcal{Z}, \mathcal{Z}^*} = 0.$$

Proof. See [85, Section 6.1] and the reference therein. \square

In Section 5.1.1 we will introduce an adjoint state associated to $(\bar{q}, \bar{\mu})$ and recover the first-order optimality conditions in *Karush-Kuhn-Tucker* (KKT) form through the Lagrangian formalism.

The result above continues to hold for the non-convex case substituting the regularity condition (2.8) with the linearized Slater's condition

$$(2.11) \quad G(\bar{q}) + G'(\bar{q})(\tilde{q} - \bar{q}) \in \text{int } \mathcal{K}.$$

However, in this case the first-order conditions are necessary but not sufficient for local optimality and second-order conditions must be postulated.

The quest for second order conditions in parabolic optimal control problem with state constraints is an open problem, as we have discussed in the introduction. Hereafter we give a brief account, coming back to the matter in Section 5.1.2.

In presence of state constraints, a cone of critical direction $C_{\bar{q}}$ is introduced which, as the name suggests, includes the direction $p \in \mathcal{Q}$ associated with \bar{q} where optimality is searched for. Then, the necessary second order condition is expressed by

$$(2.12a) \quad \frac{\partial \mathcal{L}}{\partial q^2}(\bar{q}, \bar{\mu})p \geq 0, \quad \forall C_{\bar{q}}$$

while the sufficient condition is

$$(2.12b) \quad \frac{\partial \mathcal{L}}{\partial q^2}(\bar{q}, \bar{\mu})p > 0, \quad \forall C_{\bar{q}} \setminus \{0\}.$$

Assuming the sufficient condition above, one derives the so-called quadratic growth condition which, for constant $\delta, \eta > 0$ and a Banach space $\tilde{\mathcal{Q}}$, reads

$$(2.13) \quad j(q) \geq j(\bar{q}) + \delta \|q - \bar{q}\|_{\tilde{\mathcal{Q}}}^2,$$

for any feasible control q such that $\|q - \bar{q}\|_{\tilde{\mathcal{Q}}} \leq \eta$.

Typically, the Banach space $\tilde{\mathcal{Q}}$ possesses a norm stronger than the one of \mathcal{Q} , where condition (2.13) holds, and it is the space where the functional $j(\cdot)$ is twice differentiable. This fact, called two-norm discrepancy, see [44], is a common issue in the context of non-convex optimal control. However, we will see in Section 5.1.2 that it can be eliminated in our case, i.e., we can work with $\tilde{\mathcal{Q}} = \mathcal{Q}$. We conclude this overview anticipating that the quadratic growth condition (2.13) will be the base for the derivation of convergence rates for the non-convex case in Section 5.3.

2.2.3 Discretization

In this section, we describe the discretization of the problem based on space-time finite elements methods. Namely, we use the discontinuous in time and continuous in space Galerkin method.

As the name suggests, this method allows discontinuities at the nodal points of the temporal discretization. In Chapter 4, we will derive the error at such nodal points, as well as in the interior of the time intervals, exploiting the fact that, for each time interval, we have an error equation which is independent from the previous time intervals. Another advantage relies on the fact that the time discrete problem admits a variational formulation making the method suitable for an Aubin-Nitsche argument.

The discontinuous in time and continuous in space Galerkin method has been firstly introduced in the context of parabolic equations in [46]. Afterwards, [57] used a backward Euler scheme for the time discretization which coincides with the method at hand in the case of piecewise constant polynomials. For a thorough discussion of the method, we refer to [84, Chapter 12] and the reference therein. The control variable is discretized implicitly by the optimality conditions using the so-called variational discretization, dating back to [41].

Time Discretization Let t_i be time points such that

$$0 = t_0 < t_1 < \dots < t_{N-1} < t_N = T$$

which define intervals $I_n = (t_{n-1}, t_n]$ of size k_n with $k := \max_n k_n$. Then, the intervals $I_n = (t_{n-1}, t_n]$, for $n = 1, \dots, N$, together with $I_0 = \{0\}$ form a partition of \bar{I} .

We impose the following technical assumptions on the time mesh which will be exploited in the derivation of error estimates for the state equation.

Assumption 2.2.1. *There exists strictly positive constants a, c, \tilde{k} such that*

$$\min_{n>0} k_n \geq ck^a, \quad \tilde{k}^{-1} \leq \frac{k_n}{k_{n+1}} \leq \tilde{k}, \quad \forall n > 0.$$

In the discontinuous Galerkin method, we seek for an approximation of the solution of the state equation (2.2) which is a polynomial with coefficients in $H_0^1(\Omega)$. We formalize this introducing, for a given positive integer r , the semi-discrete state and trial space

$$U_k^r(H_0^1(\Omega)) = \left\{ \varphi_k \in L^2(I, H_0^1(\Omega)) \mid \varphi_{k,n} = \varphi_k|_{I_n} \in \mathcal{P}_r(I_n, H_0^1(\Omega)), n = 1, \dots, N \right\}$$

with inner product and norm given by the restriction of inner product and norm of $L^2(I, L^2(\Omega))$ on I_n , that is,

$$(\cdot, \cdot)_{I_n} := \int_{I_n} (\cdot, \cdot) dt, \quad \|\cdot\|_{I_n} := \|\cdot\|_{L^2(I_n, L^2(\Omega))}.$$

Functions in $U_k^r(H_0^1(\Omega))$ can have discontinuities at the time points t_i and notation-wise we should account for this. In particular, we define right and left values at each time point as well as the corresponding jump of the functions. For a function $\varphi_k \in U_k^r(H_0^1(\Omega))$, we have

$$\varphi_{k,n}^+ := \lim_{t \rightarrow 0^+} \varphi_k(t_n + t), \quad \varphi_{k,n}^- := \lim_{t \rightarrow 0^+} \varphi_k(t_n - t) = \varphi_k(t_n), \quad [\varphi_k]_n := \varphi_{k,n}^+ - \varphi_{k,n}^-$$

and we note that the functions are continuous to the left.

We anticipate here that the bilinear form appearing in (2.3) corresponds to the standard case of the Laplace operator. Then, for $u_k, \varphi_k \in U_k^r(H_0^1(\Omega))$, we introduce the bilinear form

$$B(u_k, \varphi_k) := \sum_{n=1}^N (\partial_t u_k, \varphi_k)_{I_n} + (\nabla u_k, \nabla \varphi_k)_I + \sum_{n=2}^N ([u_k]_{n-1}, \varphi_{n-1}^+) + (u_{k,0}^+, \varphi_0^+).$$

In the subsequent analysis, we will consider approximation functions being piecewise constant polynomial, i.e., $r = 0$ with state and trial space

$$(2.14) \quad U_k := U_k^0(H_0^1(\Omega)) = \left\{ \varphi_k \in L^2(I, H_0^1(\Omega)) \text{ such that } \varphi_{k,n} = \varphi_k|_{I_n} \in \mathcal{P}_0(I_n, H_0^1(\Omega)), n = 1, \dots, N \right\}.$$

In this case, we simplify the notation and for $\varphi_k \in U_k$ we write

$$\varphi_{k,n+1} := \varphi_{k,n}^+, \quad \varphi_{k,n} := \varphi_{k,n}^-, \quad [\varphi_k]_n := \varphi_{k,n+1} - \varphi_{k,n}.$$

Then, for $u_k, \varphi_k \in U_k$, the bilinear form reduces to

$$(2.15) \quad B(u_k, \varphi_k) = (\nabla u_k, \nabla \varphi_k)_I + \sum_{n=2}^N ([u_k]_{n-1}, \varphi_{k,n}) + (u_{k,1}, \varphi_{k,1})$$

Remark 2.2.6. For the space U_k^0 on a generic V , we will use the notation $U_k(V)$.

Space Discretization We discretize the problem in space by means of conforming finite elements. To this end, we consider a family \mathcal{T}_h of subdivisions consisting of closed triangles or quadrilaterals T in dimension two and tetrahedral or hexahedral in dimension three. Such elements T are assumed to be affine

2.2. Problem setting

equivalent to their reference elements and the union $\Omega_h = \text{int}\left(\bigcup_{T \in \mathcal{T}_h} T\right)$ is such that the vertices on $\partial\Omega_h$ are located on $\partial\Omega$. We denote by h_T the diameter of the element T , set

$$h := \max_{T \in \mathcal{T}_h} h_T,$$

and we assume the family \mathcal{T}_h to be quasi-uniform and shape regular in the sense of [6].

Then, we define in the usual way the conforming finite element space $V_h \subset V$ as the space of piecewise linear function with respect to the mesh \mathcal{T}_h and we set $v_h|_{\Omega \setminus \Omega_h} \equiv 0$ for any $v_h \in V_h$.

The discrete state and trial space is defined as

$$(2.16) \quad U_{kh} := U_{kh}(V_h) = \{\varphi_{kh} \in L^2(I, V_h) \mid \varphi_{kh}|_{I_n} \in \mathcal{P}_0(I_n, V_h), n = 1, \dots, N\},$$

and the bilinear form is given as in (2.15) with $u_{kh}, \varphi_{kh} \in U_{kh}$, namely

$$(2.17) \quad B(u_{kh}(q), \varphi_{kh}) := (\nabla u_{kh}, \nabla \varphi_{kh})_I + \sum_{n=2}^N ([u_{kh}]_{n-1}, \varphi_{kh,n}) + (u_{kh,1}, \varphi_{kh,1}).$$

Once the finite elements space V_h is defined, we introduce the inverse of the discrete Laplacian

$$-\Delta_h : H^{-1}(\Omega) \rightarrow V_h,$$

which associates to any $f \in H^{-1}(\Omega)$ an element $v_h \in V_h$ given by

$$(\nabla v_h, \nabla \varphi_h) = f(\varphi_h), \quad \forall \varphi_h \in V_h.$$

Further, we will often use the discrete semi-norm

$$|\cdot|_{-s,h} = (-\Delta_h^{-s}, \cdot)^{1/2}, \quad s = 1, 2,$$

which is equivalent to the continuous semi-norm modulo a small constant, see [84, Lemma 5.3].

Control Discretization The control variable is discretized implicitly by the optimality conditions via the variational discretization approach introduced in [41] for elliptic problems and extended in [22] for the parabolic case. In view of Q_{ad} given by box-constraints, denoting with $P_{Q_{ad}}$ the usual pointwise projection onto Q_{ad} , with \bar{z}_k, \bar{z}_{kh} the semi-discrete and discrete adjoint state arising from the corresponding KKT-system, respectively, we have

$$(2.18) \quad \bar{q}_k = P_{Q_{ad}}\left(-\alpha^{-1}\bar{z}_k\right), \quad \bar{q}_{kh} = P_{Q_{ad}}\left(-\alpha^{-1}\bar{z}_{kh}\right).$$

In view of the regularity of \bar{z}_k, \bar{z}_{kh} from Section 3.3 and the presence of box-constraints, this means that

$$\bar{q}_k|_{I_n}, \bar{q}_{kh}|_{I_n} \in \mathcal{P}_0(I_n, H^1(\Omega)).$$

We don't investigate other types of control discretization because the aim of this thesis is to analyze state constraints and different control discretizations would only add more technicalities not related with our main goal.

Other types of control discretization in the context of parabolic optimal control problems have been analyzed in [61] for a setting with state constraints, and in [63, 68] for control constraints only.

2.3 Examples

In this section, we state the problems under consideration making concrete what we have presented in the previous sections.

The first problem has been analyzed by the author of this thesis in [56] and it is characterized by a linear state equation and a point-wise in time constraints on weighted-mean values of the gradient of the state variable. The error analysis for the state equation will be done in Section 4.3 while the convergence analysis for the optimal control problem will be given in Section 5.2.1.

The other problem, considered in [55] by the author of this work, presents a semi-linear state equation and point-wise in time constraints on weighted-mean value of the state variable. In Section 4.4 we will derive error estimates for the state equation and in Section 5.3 the convergence rate of the optimization problem.

We now introduce, for $q_{\min} < q_{\max}$, the set of admissible controls denoted by

$$(2.19) \quad Q_{\text{ad}} = \left\{ q \in L^2(I, \mathbb{R}^m) \mid q_{\min} \leq q(t) \leq q_{\max} \right\}.$$

2.3.1 First-order state constraint with linear state equation

For given $q_i \in L^2(I)$, $g_i \in H_0^1(\Omega)$, $i = 1, \dots, m$, initial data $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ and prescribed temperature profile $u_d \in L^2(I, L^2(\Omega))$, we consider the problem

$$(2.20a) \quad \text{Minimize } J(q, u) = \frac{1}{2} \|u - u_d\|_I^2 + \frac{\alpha}{2} \|q\|_{L^2(I)}^2,$$

where the state $u(t, x)$ and the control $q(t) = (q_i(t))_{i=1}^m$ are coupled by the linear PDE

$$(2.20b) \quad \begin{aligned} \partial_t u(t, x) - \Delta u(t, x) &= \sum_{i=1}^m q_i(t) g_i(x) && \text{in } I \times \Omega, \\ u(t, x) &= 0 && \text{on } I \times \partial\Omega, \\ u(0, x) &= u_0 && \text{in } \{0\} \times \Omega, \end{aligned}$$

subject to control constraints

$$(2.20c) \quad q \in Q_{\text{ad}} \quad \text{a.e. in } I,$$

and point-wise in time gradient state constraint

$$(2.20d) \quad F(u)(t) := \int_{\Omega} |\nabla u(t, x)|^2 \omega(x) dx \leq b \quad \forall t \in [0, T],$$

for weighting function $\omega \in L^\infty(\Omega)$.

For $u, \varphi \in W(0, T)$ we define the bilinear form

$$(2.21) \quad b(u, \varphi) = (\partial_t u, \varphi)_I + (\nabla u, \nabla \varphi)_I + (u(0), \varphi(0))$$

and the weak formulation of the problem reads: for given $q \in L^2(I, \mathbb{R}^m)$ and $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ find $u \in W(0, T)$ such that

$$(2.22) \quad b(u, \varphi) = (qg, \varphi)_I + (u_0, \varphi(0)), \quad \forall \varphi \in W(0, T).$$

2.3. Examples

Actually, the state variable u possesses additional regularity but we postpone this point to the next chapter.

In a similar fashion, we define semi-discrete and discrete state equation corresponding to the time and space discretization, respectively, again with data q, g and u_0 as in (2.22): find $u_k = u_k(q) \in U_k$ such that

$$(2.23) \quad B(u_k, \varphi_k) = (qg, \varphi_k)_I + (u_0, \varphi_{k,1}), \quad \forall \varphi_k \in U_k;$$

find $u_{kh} = u_{kh}(q) \in U_{kh}$

$$(2.24) \quad B(u_{kh}, \varphi_{kh}) = (qg, \varphi_{kh})_I + (u_0, \varphi_{kh,1}), \quad \forall \varphi_{kh} \in U_{kh}.$$

We remark one more time that the control is not discretized. The semi-discrete and discrete optimal control problem read

$$(2.25) \quad \begin{aligned} & \underset{(q, u_k) \in Q_{ad} \times U_k}{\text{Minimize}} \quad J(q, u_k) = \frac{1}{2} \|u_k - u_d\|_I^2 + \frac{\alpha}{2} \|q\|_{L^2(I)}^2 \\ & \text{subject to (2.23) and} \\ & F(u_k)|_{I_n} \leq b, \quad n = 1, \dots, N, \end{aligned}$$

and

$$(2.26) \quad \begin{aligned} & \underset{(q, u_{kh}) \in Q_{ad} \times U_{kh}}{\text{Minimize}} \quad J(q, u_{kh}) = \frac{1}{2} \|u_{kh} - u_d\|_I^2 + \frac{\alpha}{2} \|q\|_{L^2(I)}^2 \\ & \text{subject to (2.24) and} \\ & F(u_{kh})|_{I_n} \leq b, \quad n = 1, \dots, N, \end{aligned}$$

respectively.

2.3.2 Zero-order state constraint with semi-linear state equation

In a first step, we pose some assumptions on the semi-linear term which can be regarded as classical in the context of non-convex optimization, see [85, Assumption 5.6].

Assumption 2.3.1. *The nonlinear term $d(t, x, u): I \times \Omega \times \mathbb{R}$ is assumed to satisfy the following:*

- (i) *For all $u \in \mathbb{R}$, the nonlinearity is measurable with respect to $(t, x) \in I \times \Omega$. Further, for almost every $(t, x) \in I \times \Omega$ it is four time continuously differentiable with respect to u .*
- (ii) *For $u = 0$, there is $c > 0$ such that $d(t, x, 0)$ satisfies, together with its derivatives up to order two, the boundedness condition*

$$\|d(\cdot, \cdot, 0)\|_{L^\infty(I \times \Omega)} + \|\partial_u d(\cdot, \cdot, 0)\|_{L^\infty(I \times \Omega)} + \|\partial_u^2 d(\cdot, \cdot, 0)\|_{L^\infty(I \times \Omega)} \leq c.$$

Further, each of these satisfy a local Lipschitz condition with respect to u , i.e., for any $M > 0$ there exist a constant $L(M) > 0$ such that for any $|u_j| \leq M$ $j = 1, 2$ there holds

$$\|\partial_u^i d(\cdot, \cdot, u_1) - \partial_u^i d(\cdot, \cdot, u_2)\|_{L^\infty(I \times \Omega)} \leq L(M)|u_1 - u_2|,$$

for every $i = 0, 1, 2$.

(iii) For all $u \in \mathbb{R}$ and for almost every $(t, x) \in I \times \Omega$, there holds the monotonicity condition

$$\partial_u d(t, x, u) \geq 0.$$

When no confusion arises, we shorten the notation from $d(\cdot, \cdot, u)$ to $d(u)$.

Remark 2.3.1. Comparing the first assumption (i) with the corresponding one in [85, Assumption 5.6], we note that we are assuming $d(\cdot, \cdot, u)$ to be four times differentiable rather than twice. This is the regularity that we will need in Section 4.2.2 to obtain error estimates for the semi-linear case.

We now state the non-convex optimal control problem. For $q_i \in L^2(I)$, $g_i \in L^\infty(\Omega)$, $i = 1, \dots, m$, initial data $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ and $u_d \in L^2(I, L^2(\Omega))$, we consider the problem

$$(2.27a) \quad \text{Minimize } J(q, u) = \frac{1}{2} \|u - u_d\|_I^2 + \frac{\alpha}{2} \|q\|_{L^2(I)}^2,$$

where the state $u(t, x)$ and the control $q(t) = (q_i)_{i=1}^m$ are coupled by the semi-linear PDE

$$(2.27b) \quad \begin{aligned} \partial_t u(t, x) - \Delta u(t, x) + d(t, x, u(t, x)) &= \sum_{i=1}^m q_i(t) g_i(x) && \text{in } I \times \Omega, \\ u(t, x) &= 0 && \text{on } I \times \partial\Omega, \\ u(0, x) &= u_0 && \text{in } \{0\} \times \Omega, \end{aligned}$$

subject to control constraints

$$(2.27c) \quad q \in Q_{\text{ad}} \quad \text{a.e. in } I$$

and state constraints

$$(2.27d) \quad F(u)(t) = \int_{\Omega} u(t, x) \omega(x) dx \leq b \quad \forall t \in [0, T],$$

for weighting function $\omega \in L^\infty(\Omega)$.

With the bilinear form defined as in (2.21), the weak form of the state equation reads: for given $q \in L^2(I, \mathbb{R}^m)$ and $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ find $u \in W(0, T)$ such that

$$(2.28) \quad b(u, \varphi) + (d(\cdot, \cdot, u), \varphi)_I = (qg, \varphi)_I + (u_0, \varphi(0)), \quad \forall \varphi \in W(0, T).$$

Also in this case the solution u possesses additional regularity.

With q, g and u_0 as above, the semi-discrete state equation reads: find $u_k = u_k(q) \in U_k$ such that

$$(2.29) \quad B(u_k, \varphi_k) + (d(\cdot, \cdot, u_k), \varphi_k)_I = (qg, \varphi_k)_I + (u_0, \varphi_{k,1}), \quad \forall \varphi_k \in U_k.$$

The discrete state equation consists in finding $u_{kh} = u_{kh}(q) \in U_{kh}$

$$(2.30) \quad B(u_{kh}, \varphi_{kh}) + (d(\cdot, \cdot, u_{kh}), \varphi_{kh})_I = (qg, \varphi_{kh})_I + (u_0, \varphi_{kh,1}), \quad \forall \varphi_{kh} \in U_{kh}.$$

2.3. Examples

Then, the semi-discrete and discrete optimal control problem read

$$\begin{aligned}
 (2.31) \quad & \underset{(q, u_k) \in Q_{ad} \times U_k}{\text{Minimize}} \quad J(q, u_k) = \frac{1}{2} \|u_k - u_d\|_I^2 + \frac{\alpha}{2} \|q\|_{L^2(I)}^2 \\
 & \text{subject to (2.29) and} \\
 & F(u_k)|_{I_n} \leq b, \quad n = 1, \dots, N,
 \end{aligned}$$

and

$$\begin{aligned}
 (2.32) \quad & \underset{(q, u_{kh}) \in Q_{ad} \times U_{kh}}{\text{Minimize}} \quad J(q, u_{kh}) = \frac{1}{2} \|u_{kh} - u_d\|_I^2 + \frac{\alpha}{2} \|q\|_{L^2(I)}^2 \\
 & \text{subject to (2.30) and} \\
 & F(u_{kh})|_{I_n} \leq b, \quad n = 1, \dots, N,
 \end{aligned}$$

respectively.

3. Parabolic partial differential equations

This chapter is concerned with the regularity of the solutions of the parabolic PDEs introduced in Section 2.3, and of further associated problems which will appear in the following chapters. Most of the results presented here are well-established in the literature. Hence we confine the exposition to a formal and precise recollection to avoid tedious repetitions. However, for a complete treatment of the subject, we will always disclose the methods employed in the proof of these results and where they can be found in the literature.

In Section 3.1, we deal with linear parabolic equations. The main reference for this part are the classical monographs [30, 52, 53]. The extension to semi-linear differential equations is performed in Section 3.2, using results from the seminal paper [7] based on the classical truncation procedure of Stampacchia. Section 3.3 is devoted to the study of differential equations having a measure as data. Such problems will naturally appear when deriving the optimality conditions for the problems under consideration. In the final Section 3.4, we review the regularity of several functionals appearing in the semi-linear problem.

3.1 Well-posedness and regularity of solutions for linear equations

We start the analysis considering the regularity of the linear differential equation defined in (2.20b). As already anticipated in Section 2.2.1, it is important that the resulting regularity allows the embedding of the state space into a suitable space of continuous functions.

Proposition 3.1.1. *For $q_i \in L^2(I, \mathbb{R})$, $g_i \in L^2(\Omega)$, with $i = 1, \dots, m$, and $u_0 \in H_0^1(\Omega)$, there exists a unique solution*

$$(3.1) \quad u \in U := L^2(I, H_0^1 \cap H^2(\Omega)) \cap H^1(I, L^2(\Omega))$$

of the equation

$$(3.2) \quad b(u, \varphi) = (qg, \varphi)_I + (u_0 + \varphi(0))$$

for any $\varphi \in W(0, T)$, where $b(\cdot, \cdot)$ is defined as in (2.21).

Proof. We refer to [53, Section 5, Theorem 5.3] for the existence of the solution in the spaces $L^2(I, H_0^1(\Omega) \cap H^2(\Omega)) \cap H^1(I, L^2(\Omega))$. It is obtained, in Lions and Magenes words, using *basic inequalities* in order to construct an isomorphism between Hilbert spaces.

The regularity $L^\infty(I, H_0^1(\Omega))$ is obtained using the Galerkin approximation, see [30, Chapter 7, Theorem 5], where same state space U as here is obtained in a more modern notation with respect to [53]. \square

We observe that the gradient state constraint (2.20d)

$$(|\nabla u|^2, \omega) = F: U \rightarrow C(\bar{I})$$

is well posed as there holds the embedding

$$(3.3) \quad U \hookrightarrow C(\bar{I}, H_0^1(\Omega)),$$

see [52, Chapter 1, Theorem 3.1]. The regularity $W(0, T)$ would have lead to the embedding into the space $C(\bar{I}, L^2(\Omega))$ which is not enough for the treatment of (2.20d). This justifies the need for additional regularity.

Remark 3.1.2. *In Proposition 3.1.1, we have assumed the minimal regularity on the data to ensure the embedding above. For the rest of this work in relation with Problem 2.20, we will assume that*

$$g \in H_0^1(\Omega) \text{ and } u_0 \in \dot{H}^3(\Omega)$$

in order to fully exploit the approximation properties of the time discretization.

Remark 3.1.3. *The existence and regularity of the solutions of the semi-discrete and discrete state equation (2.23) and (2.24), respectively, follows from standard arguments of elliptic theory. We refer, e.g., to [84, Chapter 12] and the reference therein.*

Denoting with $u(q)$ and $u_k(q)$ the solutions of (2.22) and (2.23) associated with

$q \in Q_{ad}$, respectively, we note that $u(q)$ satisfies also the semi-discrete state equation

$$B(u(q), \varphi_k) = (qg, \varphi_k) + (u_0, \varphi_{k,1}), \quad \forall \varphi_k \in U_k.$$

This is readily seen noting that in particular $u \in C(\bar{I}, L^2(\Omega))$.

As a consequence, there holds the following Galerkin orthogonality relation

$$(3.4) \quad B(u(q) - u_k(q), \varphi_k) = 0, \quad \forall \varphi_k \in U_k.$$

We conclude this section with a regularity result for the adjoint (uncontrolled) counterpart of (2.20b) which will be used in the forthcoming chapter. The second part of the following proposition corresponds to [56, Lemma 4.2] of the author of this thesis.

Proposition 3.1.4. *For a given $w_T \in H^{-1}(\Omega)$ there exists a unique solution*

$$(3.5) \quad w \in W := L^2(I, L^2(\Omega)) \cap H^1(I, (\dot{H}^2)^*)$$

of the equation

$$(3.6) \quad -(\varphi, \partial_t w)_I + (\nabla \varphi, \nabla w)_I = 0, \quad w(T) = w_T$$

for any $\varphi \in L^2(I, \dot{H}^2(\Omega)) \cap H^1(I, L^2(\Omega))$. Further, there holds the stability estimate

$$(3.7) \quad \|w\|_I + \max_{t \in I} \|w(t)\|_{H^{-1}(\Omega)} \leq C \|w_T\|_{H^{-1}(\Omega)}$$

Proof. The existence and regularity is obtained using the transposition of the isomorphism of the primal problem in [53, Chapter 4, Section 8]. To obtain (3.7) we test (3.6) with $\varphi = -\Delta^{-1}w$

$$(3.8) \quad (\Delta^{-1}w, \partial_t w)_I - (\nabla \Delta^{-1}w, \nabla w)_I = 0.$$

We note that (3.8) holds also point-wise almost everywhere on I . Then

$$(3.9) \quad (\Delta^{-1}w(t), \partial_t w(t)) - (\nabla \Delta^{-1}w(t), \nabla w(t)) = 0, \quad \text{for a.e. } t \in I.$$

Using the relation $\partial_t w = -\Delta w$, we reformulate the first term obtaining

$$(\Delta^{-1}w(t), \partial_t w(t)) = -\|w(t)\|^2.$$

For the second, we exploit the definition of $-\Delta^{-1}$ via its weak form, i.e., $(\nabla \cdot, \nabla \Delta^{-1} \cdot) = -(\cdot, \cdot)$, together with (3.9),

$$\begin{aligned} -(\nabla \Delta^{-1}w(t), \nabla w(t)) &= -(\Delta^{-1}w(t), \partial_t w(t)) \\ &= (\nabla \Delta^{-1}w(t), \nabla \Delta^{-1} \partial_t w(t)). \end{aligned}$$

Then, observing that the time derivative interchanges with ∇ and Δ^{-1} , we have

$$(\nabla \Delta^{-1}w(t), \nabla \Delta^{-1} \partial_t w(t)) = \frac{1}{2} \frac{d}{dt} \|\nabla \Delta^{-1}w(t)\|^2.$$

Thus, it follows from (3.9) that

$$(3.10) \quad \frac{d}{dt} \|\nabla \Delta^{-1}w(t)\|^2 = 2\|w(t)\|^2.$$

Integrating (3.10) over (t, T) and defining $\eta(t) = \|\nabla \Delta^{-1} w(t)\|^2, \psi(t) = \|w(t)\|^2$, we obtain

$$\eta(t) + 2 \int_t^T \psi(s) ds = \eta(T).$$

Noting that both η and ψ are nonnegative, the claim is shown. \square

We will come back to (3.6) in Section 4.2.1 where we will derive time-weighted estimate for its solution.

We observe that, considering the spatial part involved in the definition of W as an interpolation couple, there holds

$$[L^2(\Omega), (\dot{H}^2(\Omega))^*]_{\frac{1}{2}} = H^{-1}(\Omega),$$

see [52, Chapter 1, Theorem 12.5].

As a consequence, the space W is embedded into a space of continuous functions, namely,

$$W \hookrightarrow C(\bar{I}, H^{-1}(\Omega)),$$

see [52, Chapter 1, Theorem 3.1], as one expects from the problem at hand.

Remark 3.1.5. *The interpolation result in [52] is given for the space $(H^2(\Omega))^*$. However, a study of the proof, based on the reiteration principle, reveals that it continues to hold also for $(\dot{H}^2(\Omega))^*$.*

3.2 Semi-linear differential equations

We move our attention to the semi-linear state equation introduced in (2.27b). Firstly, we infer the minimal regularity necessary to allow the embedding of the state space in a suitable space of continuous functions. Then, we obtain additional regularity which will come into play to ensure the Lipschitz continuity of the resulting control-to-state map. We now require the strong convergence of the sequence of state variables to guarantee the convergence of the Nemytskii operator representing the semi-linear term. Going back to Theorem 2.2.4, this is indeed what we will need in Section 5.1.1 to obtain the existence of a local solution in the non-convex case.

In the second part of this section, we focus on the properties of the solutions of the semi-discrete and discrete state equation which will be important later in the analysis of the discretization.

Proposition 3.2.1. *Under Assumption 2.3.1, $q_i \in L^2(I, \mathbb{R}), g_i \in L^\infty(\Omega)$, with $i = 1, \dots, m$, and $u_0 \in H_0^1(\Omega) \cap C(\bar{\Omega})$, there exists a unique solution*

$$(3.11) \quad u \in W(0, T) \cap C(\bar{I} \times \bar{\Omega})$$

of the equation

$$(3.12) \quad b(u, \varphi) + (d(t, x, u), \varphi)_I = (qg, \varphi)_I + (u_0, \varphi(0))$$

for any $\varphi \in W(0, T)$, where $b(\cdot, \cdot)$ is defined as in (2.21).

Further, for $r > n/2 + 1$, if $\{q_k\}_{k=1}^\infty$ is a sequence converging weakly in $L^{\tilde{r}}(I, \mathbb{R}^m)$ to q , with $\tilde{r} = \max\{r, 2\}$, then the sequence $\{u(q_k)\}_{k=1}^\infty$ converges to the solution $u = u(q)$ strongly in $C(\bar{I} \times \bar{\Omega})$.

Proof. The existence and regularity in $W(0, T)$ of the solution, in the context of optimal control problem, has been firstly obtained in the seminal paper [7, Theorem 5.1]. The idea behind is based on a method originally introduced in [82, 83], see also [85, Section 7.2.2], for semi-linear elliptic equations making use of cut-off functions.

The continuity of the solution is a consequence of the assumptions on the semi-linear term $d(\cdot, \cdot, u)$. Indeed, by moving this term in the right-hand side of (3.12) and exploiting its boundedness and Lipschitz continuity, one is reconduct to the study of a linear equation.

The formulation of these results in our setting has been obtained in [20, Theorem 1]. Further, the authors have obtained the strong convergence of the sequence of state variables using the concept of maximal parabolic regularity, see [38].

An important feature to note is that we do not have any limitation on the dimension of the domain due to the form of the right-hand side of (3.12) where the spatial part g is fixed. Then, the resulting state variable u associated with $q_i, i = 1, \dots, m$, is continuous. When the control variable acts in time and space, to ensure the continuity one either works in the one-dimensional setting or needs $q \in L^r(I \times \Omega)$ with $r > n/2 + 1$, compare with [8, Theorem 7.2] and [17, Theorem 4.1]. \square

Also for the semi-linear case, the state constraint (2.27d)

$$(u, \omega) = F: W(0, T) \rightarrow C(\bar{I})$$

is well-posed due to the embedding

$$(3.13) \quad W(0, T) \hookrightarrow C(\bar{I}, L^2(\Omega)),$$

see [52, Chapter 1, Theorem 3.1].

In a next step, we see that the solution u of the semi-linear state equation exhibits additional regularity. This is required to ensure Lipschitz continuity of the control-to-state map in Section 3.4.

Proposition 3.2.2. *Under Assumption 2.3.1, for the solution u of (2.27b) there holds the additional regularity*

$$(3.14) \quad u \in L^2(I, H_0^1(\Omega) \cap H^2(\Omega)),$$

and the following stability estimates hold

$$(3.15) \quad \begin{aligned} \|u\|_{L^\infty(I \times \Omega)} &\leq c(\|qg\|_{L^\infty(I \times \Omega)} + \|u_0\|_{L^\infty(\Omega)} + \|d(\cdot, \cdot, 0)\|_{L^\infty(I \times \Omega)}), \\ \|u\|_{L^2(I, H_0^1(\Omega))} + \|u\|_{L^2(I, H^2(\Omega))} + \|u\|_{L^\infty(I, H_0^1(\Omega))} + \|\partial_t u\|_I \\ &\leq c(\|qg\|_I + \|u_0\|_{H_0^1(\Omega)} + \|d(\cdot, \cdot, 0)\|_I). \end{aligned}$$

Proof. The first relation is a direct consequence of Proposition 3.2.1. For the second stability estimate we refer to [68, Proposition 2.1]. The idea consists in moving the semi-linear term to the right-hand side where one exploits the Lipschitz continuity in $L^2(I, L^2(\Omega))$ of $d(\cdot, \cdot, u)$ in combination with Assumption 2.3.1(ii) and the boundedness of u . Then, one is reconducted to a linear equation where Proposition 3.1.1 is valid. \square

3.2. Semi-linear differential equations

In the rest of the thesis, we denote the state space associated with problem (2.27b) by

$$(3.16) \quad U := \{u \in L^2(I, H_0^1(\Omega) \cap H_0^2(\Omega)) \cap C(\bar{I} \times \bar{\Omega}) \cap H^1(I, L^2(\Omega))\}.$$

We investigate now the regularity of the solution of the semi-discrete and discrete state equation (2.29) and (2.30), respectively. In particular, we are interested in the boundedness in $L^\infty(I \times \Omega)$, independently from the discretization parameters k, h , on which we will rely in the next chapters when deriving convergence rates. For the existence and uniqueness of the solution of (2.29) and (2.30), the same considerations of Remark 3.1.3 are true.

Proposition 3.2.3. *Under Assumption 2.3.1, there exists a constant C independent from the mesh size k such that for the solution $u_k \in U_k$ of (2.29) it holds*

$$(3.17) \quad \|u_k\|_{L^\infty(I \times \Omega)} \leq C(\|qg\|_{L^r(I \times \Omega)} + \|u_0\|_{L^\infty(\Omega)} + \|d(\cdot, \cdot, 0)\|_{L^r(I \times \Omega)})$$

for every $r > 2$. Further, there holds the following stability estimates

$$(3.18) \quad \|u_k\|_{L^\infty(I, H_0^1(\Omega))} \leq C(\|qg\|_I + \|u_0\|_{H_0^1(\Omega)} + \|d(\cdot, \cdot, 0)\|_I)$$

Proof. The boundedness in $L^\infty(I \times \Omega)$ follows from the application of the method of truncation of Stampacchia [83]. In brief, the semi-discrete state equation (2.29) is tested with a truncation of the solution u of the continuous state equation (3.12). Then, showing that this truncation vanishes almost everywhere, one infers the desired boundedness for u_k . The monotonicity of the semi-linear term $d(\cdot, \cdot, u)$ plays an important role to achieve this result. The reader is refer to [68, Theorem 3.1] where the claim is shown, see also [85, Theorem 4.5 and Lemma 7.5]. The stability estimate (3.18) has been shown in [68, Theorem 3.2]. \square

For the solution of (2.30) there holds similar conclusions.

Proposition 3.2.4. *Under Assumption 2.3.1, there exists a constant C independent of k and h such that*

$$(3.19) \quad \|u_{kh}\|_{L^\infty(I \times \Omega)} \leq C(\|qg\|_{L^r(I \times \Omega)} + \|P_h u_0\|_{L^\infty(\Omega)} + \|d(\cdot, \cdot, 0)\|_{L^r(I \times \Omega)}),$$

for every $r > 2$, where $P_h: H_0^1(\Omega) \rightarrow V_h$ is the usual L^2 -projection in space. Further, it holds the stability estimate

$$(3.20) \quad \|u_k\|_{L^\infty(I, H_0^1(\Omega))} \leq C(\|qg\|_I + \|P_h u_0\|_{H_0^1(\Omega)} + \|d(\cdot, \cdot, 0)\|_I).$$

Proof. See [68, Theorem 4.2]. \square

Remark 3.2.5. *For the rest of this work, when dealing with problem (2.27) we will assume $u_0 \in H^2(\Omega)$ in order to use results from [61, 62] to fully exploit the approximation property of the discontinuous Galerkin method.*

3.3 Differential equations with measure as data

One of the features of state-constrained optimal control problems is the presence of a measure in the differential equation defining the adjoint variable in the KKT-optimality system. A study of the regularity of this equation with respect to this measure is therefore fundamental in view of the convergence analysis.

The adjoint equations associated with constraints (2.20d) and (2.27d) have already been analyzed in the literature with different techniques. The former in [59], while the latter in several publications [5, 7, 48]. Basically, the proofs differ in the calculation of the adjoint of the solution operator of the PDE, by means of an integral representation, that is, using Green's functions, or by using the method of transposition of [53]. In all cases, their final findings agree and, in particular, an integral state constraint point-wise in time lead to a positive Borel measure in $I = [0, T]$ concentrated in the points where the state constraint is active.

Considering the importance of this matter in this work, and the relatively specific subject, in the following we review the method employed in [59], based on the transposition method, to deduce the regularity of the adjoint variable associated with the gradient constraint (2.20d). The constraints (2.27d) can be analyzed with the same technique; roughly speaking, scaling opportunely the spaces involved, see also [7, Theorem 6.4].

In a first step, one uses [53, Theorem 1.1] to deduce that

$$(\partial_t + \mathcal{A}): W := L^2(I, H_0^1(\Omega) \cap H^2(\Omega)) \cap H^1(I, L^2(\Omega)) \rightarrow L^2(I \times \Omega)$$

is an isomorphism. This implies that for every $L \in W^*$ there exists a unique $z \in L^2(I, L^2(\Omega))$ satisfying

$$L(\varphi) = (\partial_t \varphi + \mathcal{A}\varphi, z)_I$$

for all $\varphi \in W$. By means of the embedding (3.3), we clearly have

$$W \hookrightarrow C(\bar{I}, H_0^1(\Omega))$$

and, therefore, for a given $\nu \in C(\bar{I}, H_0^1(\Omega))^*$, a linear and bounded functional on W is built by

$$L(\varphi) = \langle \varphi, \nu \rangle,$$

with $\langle \cdot, \cdot \rangle$ being the pairing between $C(\bar{I}, H_0^1(\Omega))$ and its dual. In particular, this means that, for $\nu \in C(\bar{I}, H_0^1(\Omega))^*$ given, there exists a unique solution $z \in L^2(I, L^2(\Omega))$ of the differential equation

$$(3.21) \quad (\partial_t \varphi + \mathcal{A}\varphi, z)_I = \langle \varphi, \nu \rangle.$$

Remark 3.3.1. *The derivation above corresponds to [59, Lemma 3]. For an easy comparison, we provide the changes adopted in the notation*

$$z^* = \nu, \quad w^* = z, \quad \ell = L, \quad {}_0^0 W(0, T; H^2(\Omega), L^2(\Omega)) = W.$$

Further, we note that in [59] there is an observation of the state variable at the final time $t = T$ in the cost functional. This leads to the presence of an additional data in the problem, namely h^* , which does not appear in our setting.

Then, for a generic Hilbert space H , we introduce the Banach spaces

$$\begin{aligned} BV(I, H) &:= \{v: \bar{I} \rightarrow H \text{ of bounded variation} \mid v \text{ is right continuous on } I\} \\ NBV_0(I, H) &:= \{v \in BV(I, H) \mid v(T) = 0, v \text{ continuous at } t = 0, T\}. \end{aligned}$$

Using results on functions of bounded variations, see [58], in [59, Theorem 1] it is shown that the solution z of (3.21) satisfies, for given $v_\nu \in NBV_0(I, H^{-1}(\Omega))$, the integral equation

$$(3.22) \quad z(t) + \int_t^T \mathcal{A}^* z(\tau) d\tau = -v_\nu(t), \quad \forall t \in [0, T],$$

and exhibits the regularity $z \in NBV_0(I, (H^2 \cap H_0^1(\Omega))^*)$. In [59, Appendix 1] it is shown that this additional regularity implies $z \in L^\infty(I, H^{-1}(\Omega))$. This passage to an integral equation is necessary in order to obtain, in a final step, a differential equation (of vector-valued distribution) with the same structure of the adjoint equation that we will encounter in Section 5.1.1. This is obtained in [59, Theorem 2] by multiplying (3.22) with $\psi \in \mathcal{D}(I)$, space of infinitely differentiable function of I with compact support on I , and integrating over I . One obtains

$$(3.23) \quad (\partial_t \psi, z)_I + (\psi, \mathcal{A}^* z)_I = \int_0^T \psi dv_\nu, \quad \forall \psi \in \mathcal{D}(I)$$

compare with [59, Equation (2.11)]. This is indeed the form of the adjoint equation that will be treated in Section 5.1.1 in relation with the constraint (2.20d). We will come back to this equation in the proof of Theorem 5.1.2 after having explicitly introduced the convex cone of non-negative continuous function in our setting. In conclusion, for a given $v_\nu \in NBV_0(I, H^{-1}(\Omega))$, there exists a unique solution of

$$(3.24) \quad z \in L^2(I, L^2(\Omega)) \cap L^\infty(I, H^{-1}(\Omega)),$$

of (3.23).

We remark again that for the constraint (2.27d), having been treated in several publications, we do not disclose the details for the derivation of the associated regularity of the adjoint variable. We refer the reader to [7, Theorem 6.4] where the following regularity is obtained

$$(3.25) \quad z \in L^2(I, H_0^1(\Omega)) \cap L^\infty(I, L^2(\Omega)).$$

This is indeed what one expects by scaling the spaces in (3.24) for the case of a constraint on the state variable and not on its gradient.

3.4 Differentiability of the functionals

In this section, we review the properties of the operators and functionals associated with the non-convex problem (2.27) in term of their differentiability.

In a first step, we introduced the control-to-state map

$$(3.26) \quad S: L^\infty(I, \mathbb{R}^m) \rightarrow W(0, T) \cap C(\bar{I} \times \bar{\Omega}),$$

associating to any given q the solution $u = u(q) = S(q)$ of (2.28) whose well-posedness is guaranteed by Proposition 3.2.1.

Once the control-to-state map has been defined, we introduce the reduced cost functional

$$(3.27) \quad j(q) := J(q, S(q)),$$

and the concatenation of S with the state constraint (2.27d)

$$(3.28) \quad G = (F \circ S): L^\infty(I, \mathbb{R}^m) \rightarrow C(\bar{I}).$$

We start with the differentiability properties of the control-to-state map-

Proposition 3.4.1. *The solution operator S of Problem 2.28 is of class C^2 from $L^\infty(I, \mathbb{R}^m)$ to $W(0, T)$. For $p \in L^\infty(I, \mathbb{R}^m)$, its first derivative*

$$v_p := S'(q)p$$

in the direction p is the solution of

$$(3.29) \quad b(v_p, \varphi) + (\partial_u d(\cdot, \cdot, u(q))v_p, \varphi)_I = (pg, \varphi)_I,$$

for all $\varphi \in W(0, T)$.

For $p_1, p_2 \in L^\infty(I, \mathbb{R}^m)$, its second derivative

$$v_{p_1 p_2} := S''(q)p_1 p_2$$

in the direction $p_1 p_2$ is the solution of

$$(3.30) \quad b(v_{p_1 p_2}, \varphi) + (\partial_u d(\cdot, \cdot, u(q))v_{p_1 p_2}, \varphi)_I = -(\partial_u^2 d(\cdot, \cdot, u(q))v_{p_1} v_{p_2}, \varphi)_I.$$

for all $\varphi \in W(0, T)$, with v_{p_1}, v_{p_2} given by (3.29).

Proof. To show the differentiability of the control to state map one considers the problem having the semilinear term $d(\cdot, \cdot, u)$ as a data in the right-hand side. Then, the corresponding solution operator is a continuous linear map and we exploit this continuity, together with $d(\cdot, \cdot, u)$ being twice differentiable, to obtain the differentiability of S by virtue of the implicit function theorem. For further details we refer to [85, Theorem 5.15], see also [16, Theorem 5.1] for an explicit use of the implicit function theorem.

For the proof of (3.29) and (3.30) we refer to [85, Theorem 5.9 and 5.16], respectively, see also [20, Equations (12)-(13)], where these equations are derived in strong form. \square

Remark 3.4.2. *We observe that for $S'(q)p$ there holds the stability estimates given in Proposition 3.2.2. Indeed, (3.29) has essentially the same structure as (2.28) thanks to the Lipschitz continuity and boundedness of $\partial_u d(\cdot, \cdot, u)$ together with the boundedness of u . In particular, we will use the stability estimate in the $L^\infty(I, L^2(\Omega))$ -norm.*

Remark 3.4.3. *It is clear that thanks to differentiability of the control-to-state map also its concatenation with the state constraint G is twice differentiable, see, e.g., [8].*

In a next step, we analyze the Lipschitz properties of the solution operator S and its first derivative.

Lemma 3.4.4. *For $p, q_1, q_2 \in Q_{ad}$, there exists a constant $C > 0$ such that*

$$(3.31a) \quad \|S(q_1) - S(q_2)\|_I \leq C\|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)},$$

$$(3.31b) \quad \|S(q_1) - S(q_2)\|_{L^\infty(I, L^2(\Omega))} \leq C\|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)},$$

$$(3.31c) \quad \|S(q_1) - S(q_2)\|_{L^\infty(I, H_0^1(\Omega))} \leq C\|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)},$$

$$(3.31d) \quad \|S'(q_1)p - S'(q_2)p\|_I \leq C\|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)}\|p\|_{L^2(I, \mathbb{R}^m)},$$

$$(3.31e) \quad \|S'(q_1)p - S'(q_2)p\|_{L^\infty(I, L^2(\Omega))} \leq C\|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)}\|p\|_{L^2(I, \mathbb{R}^m)}.$$

Proof. This result is a direct consequence of [68, Lemma 2.3], where the authors, for $q_1g, q_2g \in L^\infty(I \times \Omega)$, have shown that

$$\|S(q_1) - S(q_2)\|_I \leq c\|g(q_1 - q_2)\|_I.$$

Adapting the argument to the time dependent nature of the control variable, we deduce in our case that

$$\|S(q_1) - S(q_2)\|_I \leq c\|g\|_{L^\infty(\Omega)}\|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)}.$$

Similarly, we deduce (3.31b)-(3.31d).

For (3.31e), we consider $\xi := S'(q_1)p - S'(q_2)p$ and define $\tilde{u} := S'(q_2)p$. We note that, for any $\varphi \in W(0, T)$, ξ fulfills

$$(3.32) \quad b(\xi, \varphi) + (\partial_u d(u(q_1))\xi, \varphi)_I = -(\partial_u d(u(q_1))\tilde{u} - \partial_u d(u(q_2))\tilde{u}, \varphi)_I.$$

Then, by means of the stability estimate in $L^\infty(I, L^2(\Omega))$ for $S'(q)p$, see Remark 3.4.2, in combination with the Lipschitz continuity of $\partial_u d(\cdot, \cdot, u)$, we obtain

$$\begin{aligned} \|\xi\|_{L^\infty(I, L^2(\Omega))} &\leq c\|(\partial_u d(u(q_1)) - \partial_u d(u(q_2)))\tilde{u}\|_I \\ &\leq c\|u(q_1) - u(q_2)\|_{L^4(I \times \Omega)}\|\tilde{u}\|_{L^4(I \times \Omega)} \\ &\leq c\|u(q_1) - u(q_2)\|_{L^\infty(I, H_0^1(\Omega))}\|\tilde{u}\|_{L^\infty(I, H_0^1(\Omega))} \\ &\leq c\|q_1 - q_2\|_I\|p\|_I, \end{aligned}$$

where we used the embedding $L^\infty(I, H_0^1(\Omega)) \hookrightarrow L^4(I \times \Omega)$. \square

We now move our attention to the reduced cost functional.

Corollary 3.4.5. *The functional $j(q): L^\infty(I, \mathbb{R}^m) \rightarrow \mathbb{R}$ is of class C^2 in $L^\infty(I, \mathbb{R}^m)$ and for $q, p, p_1, p_2 \in L^\infty(I, \mathbb{R}^m)$ there holds*

$$\begin{aligned} j'(q)p &= \int_I \sum_{i=1}^m \left(\alpha q_i(t) + \int_\Omega (z_0(q)g_i(x))dx \right) p_i(t) dt, \\ j''(q)p_1p_2 &= \int_I \int_\Omega (v_{p_1}v_{p_2} - z_0(q)\partial_u^2 d(x, t, u(q))v_{p_1}v_{p_2}) dx dt + \int_I p_1^T \alpha p_2 dt, \end{aligned}$$

where $z_0(q) \in W(0, T)$ is the adjoint state associated with q and j , defined, for all $\varphi \in W(0, T)$ as the unique solution of

$$(3.33) \quad b(\varphi, z) + (\partial_u d(\cdot, \cdot, u(q))z, \varphi)_I = (u(q) - u_d, \varphi)_I,$$

and $v_{p_i}, i = 1, 2$ is defined as (3.29).

Proof. The differentiability of j follows from the differentiability of the solution operator S and the continuity of the norms involved in the definition of the cost functional.

The expression for j' and j'' is an application of the chain rule, see, [16, Theorem 5.2], with the modifications due to the time dependent nature of the control variable given in [20, Equations (15)-(16)]. \square

We conclude the chapter with the differentiability properties of the operators associated with the semi-discrete and discrete problem (2.31) and (2.32), respectively, which are similar to those just shown above. In the following, we omit the subscript from the test function for sake of readability.

Lemma 3.4.6. *The semi-discrete control-to-state map*

$$(3.34) \quad S_k: L^\infty(I, \mathbb{R}^m) \rightarrow U_k$$

is of class C^2 . For a given q with corresponding state $u_k = S_k(q)$ and a direction $p \in L^\infty(I, \mathbb{R}^m)$, its first derivative along p

$$(3.35) \quad v_{k,p} := S'_k(q)p$$

is the solution of

$$(3.36) \quad B(v_{k,p}, \varphi) + (\partial_u d(\cdot, \cdot, u_k) v_{k,p}, \varphi)_I = (pg, \varphi)$$

for all $\varphi \in U_k$. For $p_1, p_2 \in L^\infty(I, \mathbb{R}^m)$, its second derivative

$$(3.37) \quad v_{k,p_1 p_2} := S''_k p_1 p_2$$

is the solution of

$$(3.38) \quad B(v_{k,p_1 p_2}, \varphi) + (\partial_u d(\cdot, \cdot, u_k) v_{k,p_1 p_2}, \varphi)_I = -(\partial_u^2 d(\cdot, \cdot, u_k) v_{k,p_1} v_{k,p_2}, \varphi)_I$$

for all $\varphi \in U_k$.

Remark 3.4.7. *After extending the definition of the state constraint (2.27d) to the semi-discrete problem as*

$$(3.39) \quad F_k: U_k \rightarrow U_k(\mathbb{R}),$$

the twice differentiability of the solution operator S_k entails that

$$(3.40) \quad G_k := (F_k \circ S_k): L^\infty(I, \mathbb{R}^m) \rightarrow U_k(\mathbb{R})$$

is of class C^2 as well.

We now derive Lipschitz properties for S_k and its first derivative.

Lemma 3.4.8. *For $q_1, q_2, p \in Q_{ad}$, there exists a constant C independent of k such that there holds*

$$(3.41) \quad \|S_k(q_1) - S_k(q_2)\|_I \leq C \|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)},$$

$$(3.42) \quad \|S'_k(q_1)p - S'_k(q_2)p\|_I \leq C \|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)} \|p\|_{L^2(I, \mathbb{R}^m)},$$

$$(3.43) \quad \|S'_k(q_1)p - S'_k(q_2)p\|_{L^\infty(I, L^2(\Omega))} \leq C \|q_1 - q_2\|_{L^2(I, \mathbb{R}^m)} \|p\|_{L^2(I, \mathbb{R}^m)}.$$

3.4. Differentiability of the functionals

Proof. The first two relations follows from [68, Lemma 3.1]. For the third one we argue exactly as in (3.31e) observing that for $S'_k(q)p$ we have stability estimates as for S_k , see Remark 3.4.2. \square

For the discrete control-to-state map

$$(3.44) \quad S_{kh} : L^\infty(\Omega) \rightarrow U_{kh}$$

and the operator

$$(3.45) \quad G_{kh} : (F_{kh} \circ S_{kh}) \rightarrow U_{kh}(\mathbb{R})$$

where

$$(3.46) \quad F_{kh} : U_{kh} \rightarrow U_{kh}(\mathbb{R})$$

represents the state constraint (3.46) in the discrete level, there hold the same properties as for S_k and G_k . The first and second derivative of S_{kh} are defined by (3.36) and (3.38), respectively, with test functions from U_{kh} . Also the Lipschitz properties expressed in Lemma 3.4.8 are still valid with arguments now from [68, Lemma 4.1].

4. A priori error estimates for the state equation

This chapter is concerned with the derivation of a priori estimates for the error arising from the space-time discretization of the state equation.

The technique used is inspired by [57] where a duality argument is employed for the analysis of the backward Euler method, which coincides with the discontinuous Galerkin method of lowest order. This approach has later been extended in the seminal paper [29] to analyze the general $dG(r)$ -method; used in combination with the standard Galerkin method, the authors obtain estimates in the $L^2(I, L^2(\Omega))$ -norm, see also [62].

A further extension of such method has been employed in [27, 28] and [61] to obtain $L^\infty(I, L^2(\Omega))$ -norm estimates with similar order of convergences. In the former, time and space meshes are coupled, while in the latter they can be chosen independently from each other as in the case at hand. Another feature of [61], already exploited in [57] and referred as *smoothing property*, is the low-regularity required for the right-hand side of the state equation.

For the semi-linear state equation, we follow an idea of [70] which has been extended in [68], see also [11], to obtain estimates in the $L^\infty(I, L^2(\Omega))$ -norm. The semi-linear term is substituted by a linearized term bounded in $L^\infty(I \times \Omega)$. Prior to the derivation of $L^\infty(I, L^2(\Omega))$ -norm estimates, also estimates in $L^2(I, L^2(\Omega))$ are necessary to cope with the presence of such linearized term.

The scope of this chapter is twofold: from one side, we derive $L^\infty(I, H_0^1(\Omega))$ -norm estimates for the linear problem; from the other, we extend the approach of [68] to obtain estimates in $L^\infty(I, L^2(\Omega))$ for semi-linear equations. These estimates have been derived by the author of this thesis in [56] and [55], respectively. Further, we will argue how to extend these techniques to obtain $L^\infty(I, H_0^1(\Omega))$ -norm estimates for the semi-linear setting.

In particular, analyzing separately the influence of the temporal and spatial discretization, we obtain the following orders of convergence for the linear state equation (2.20b)

$$(4.1) \quad \|u - u_{kh}\|_{L^\infty(I, H_0^1(\Omega))} \leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h \right) \left(\|qg\|_{L^\infty(I, H_0^1(\Omega))} + \|u_0\|_{\dot{H}^3(\Omega)} \right)$$

and, for the semi-linear state equation (2.27b)

$$(4.2) \quad \|u - u_{kh}\|_{L^\infty(I, L^2(\Omega))} \leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \right) \cdot \left(\|qg\|_{L^\infty(I, L^2(\Omega))} + \|u_0\|_{H^2(\Omega)} + \|d(\cdot, \cdot, 0)\|_{L^\infty(I \times \Omega)} \right)$$

We now outline the structure of the chapter. In Section 4.1, we list a series of backward/forward auxiliary problems and disclose the technique behind the

error analysis. Section 4.2 is concerned with the stability analysis of these auxiliary problems which will be performed separately for the linear and semi-linear case. In Section 4.3, we derive $L^\infty(I, H_0^1(\Omega))$ -norm estimate for the linear state equation with corresponding estimates in negative norms for the auxiliary problems. Then, in Section 4.4, we obtain $L^\infty(I, L^2(\Omega))$ estimates for the semi-linear state equation. Finally, Section 4.5 is devoted to confirm our results numerically.

4.1 Auxiliary problems and the error analysis

We begin presenting a list of auxiliary problems which will be needed in the following sections. The regularity of their solutions is entailed by the discussion in Chapter 3. Therefore, without further ado we directly state the problems with corresponding data regularities.

We start with the standard backward counterpart of the uncontrolled state equation and the analogue problem defined on a truncated time interval. Both will be exploited to estimate the interpolation error in the the temporal discretization. In the following, we denote with W and \widehat{W} two spaces to be chosen later accordingly to the regularity of the data.

Problem 4.1.1. Find the solution $w \in W$ of

$$(4.3) \quad \begin{aligned} -(\varphi, \partial_t w)_I + (\nabla \varphi, \nabla w)_I &= 0, \\ w(T) &= w_T, \end{aligned}$$

for

- (a) given $w_T \in L^2(\Omega)$, for any $\varphi \in W(0, T)$ and $W := W(0, T)$;
- (b) given $w_T \in H^{-1}(\Omega)$ for any

$$\varphi \in L^2(I, H^2(\Omega)) \cap H^1(I, L^2(\Omega)),$$

and

$$W := L^2(I, L^2(\Omega)) \cap H^1(I, (\dot{H}^2)^*).$$

Problem 4.1.2. For the time interval $\widehat{I} = (0, \hat{t})$, $\hat{t} \in I_N$, find the solution $\widehat{w} \in \widehat{W}$ of

$$(4.4) \quad \begin{aligned} -(\varphi, \partial_t \widehat{w})_{\widehat{I}} + (\nabla \varphi, \nabla \widehat{w})_{\widehat{I}} &= 0, \\ \widehat{w}(\hat{t}) &= w_T, \end{aligned}$$

for

- (a) given $w_T \in L^2(\Omega)$, for any $\varphi \in W(0, \hat{t})$ and $\widehat{W} := W(0, \hat{t})$;
- (b) given $w_T \in H^{-1}(\Omega)$ for any

$$\varphi \in L^2(\widehat{I}, H^2(\Omega)) \cap H^1(\widehat{I}, L^2(\Omega)),$$

and

$$\widehat{W} := L^2(\widehat{I}, L^2(\Omega)) \cap H^1(\widehat{I}, (\dot{H}^2)^*).$$

Remark 4.1.3. The choice $\hat{t} \in I_N$ is purely arbitrary, dictated by the fact that in the analysis we focus our attention mostly on the last time interval I_N . We will comment how to handle the case of a general time interval I_n in the proof of Theorem 4.3.6.

Auxiliary problems are required also for the semi-discrete and discrete state equation. The definitions of the bilinear form $B(\cdot, \cdot)$ and the space $U_k(H^{-1}(\Omega))$ are those given in (2.15) and Remark 2.2.6, respectively.

Problem 4.1.4. For a given $w_T \in H^{-1}(\Omega)$, find the solution

$$(a) \quad w_k \in U_k(H^{-1}(\Omega)) \text{ of} \\ (4.5) \quad B(\varphi_k, w_k) = (\varphi_{k,N}, w_T), \quad \forall \varphi_k \in U_k;$$

$$(b) \quad w_{kh} \in U_{kh} \text{ of} \\ (4.6) \quad B(\varphi_{kh}, w_{kh}) = (\varphi_{kh,N}, w_T), \quad \forall \varphi_{kh} \in U_{kh}.$$

Problem 4.1.5. For a given $v_0 \in \dot{H}^2(\Omega)$, find the solution

$$(a) \quad v_k \in U_k \text{ of} \\ (4.7) \quad B(v_k, \varphi_k) = (v_0, \varphi_{k,1}), \quad \forall \varphi_k \in U_k;$$

$$(b) \quad v_{kh} \in U_{kh} \text{ of} \\ (4.8) \quad B(v_{kh}, \varphi_{kh}) = (v_0, \varphi_{kh,1}), \quad \forall \varphi_{kh} \in U_{kh}.$$

To treat the the semi-linear state equation, we require auxiliary problems similar to those defined above. The linearization is performed through the functions

$$(4.9) \quad \tilde{d} = \begin{cases} \frac{d(u(t,x)) - d(u_k(t,x))}{u(t,x) - u_k(t,x)} & \text{if } u(t,x) \neq u_k(t,x) \\ 0 & \text{else,} \end{cases}$$

and

$$(4.10) \quad \hat{d} = \begin{cases} \frac{d(u_k(t,x)) - d(u_{kh}(t,x))}{u_k(t,x) - u_{kh}(t,x)} & \text{if } u_k(t,x) \neq u_{kh}(t,x) \\ 0 & \text{else,} \end{cases}$$

see [68, 70]. We note that these functions are bounded thanks to the boundedness of Q_{ad} .

Problem 4.1.6. For a given $w_T \in L^2(\Omega)$, find the solution

$$(a) \quad w \in W(0, T) \text{ of} \\ (4.11) \quad -(\varphi, \partial_t w)_I + (\nabla \varphi, \nabla w)_I + (\varphi, \tilde{d}w)_I = 0, \quad \forall \varphi \in W(0, T), \\ w(T) = w_T;$$

$$(b) \quad \hat{w} \in W(0, \hat{t}) \text{ of} \\ (4.12) \quad -(\varphi, \partial_t \hat{w})_{\hat{I}} + (\nabla \varphi, \nabla \hat{w})_{\hat{I}} + (\varphi, \tilde{d}\hat{w})_{\hat{I}} = 0, \quad \forall \varphi \in W(0, T), \\ w(\hat{t}) = w_T.$$

Problem 4.1.7. For a given $w_T \in L^2(\Omega)$, find the solution

(a) $w_k \in U_k$ of

$$(4.13) \quad B(\varphi_k, w_k) + (\varphi_k, \hat{d}w_k)_I = (\varphi_k, w_T), \quad \forall \varphi_k \in U_k;$$

(b) $w_{kh} \in U_{kh}$ of

$$(4.14) \quad B(\varphi_{kh}, w_{kh}) + (\varphi_{kh}, \hat{d}w_{kh})_I = (\varphi_{kh}, w_T), \quad \forall \varphi_{kh} \in U_{kh}.$$

Problem 4.1.8. For a given $v_0 \in \dot{H}^2(\Omega)$, find the solution

(a) $v_k \in U_k$ of

$$(4.15) \quad B(v_k, \varphi_k) + (\hat{d}v_k, \varphi_k)_I = (v_0, \varphi_{k,1}), \quad \forall \varphi_k \in U_k;$$

(b) $v_{kh} \in U_{kh}$ of

$$(4.16) \quad B(v_{kh}, \varphi_{kh}) + (\hat{d}v_{kh}, \varphi_{kh})_I = (v_0, \varphi_{kh,1}), \quad \forall \varphi_{kh} \in U_{kh}.$$

We observe that similar to (3.4), the solutions of the auxiliary linearized problems satisfy the following relations

$$(4.17) \quad B(\varphi_{kh}, w_k - w_{kh}) = -(\varphi_{kh}, (w_k - w_{kh})\hat{d})_I, \quad \forall \varphi_{kh} \in U_{kh},$$

$$(4.18) \quad B(v_k - v_{kh}, \varphi_{kh}) = -((v_k - v_{kh})\hat{d}, \varphi_{kh})_I, \quad \forall \varphi_{kh} \in U_{kh},$$

while for the linear case these relations display the classical Galerkin orthogonality. We anticipate here that one of the main difference between the linear and semi-linear case relies indeed in these relation and we can already see how $L^2(I, L^2(\Omega))$ -norm estimates for the error in the dual variables will come into play for the semi-linear case.

Remark 4.1.9. In the following sections, we will also need Problem 4.1.7(a) defined through the linearization (4.9). The corresponding equation is (4.13) with \tilde{d} in place of \hat{d} . In this case, there holds the relation

$$(4.19) \quad B(\varphi_k, w - w_k) = -(\varphi_k, (w - w_k)\tilde{d})_I, \quad \forall \varphi_k \in U_k.$$

With the auxiliary problems at hand, we unveil now the approach which will be used in the rest of this section for the derivation of error estimates. The discretization error is split into its temporal and spatial part

$$\begin{aligned} e &= u - u_{kh} \\ &= u - u_k + u_k - u_{kh} \\ &= e_k + e_h. \end{aligned}$$

Then, at any level of discretization, the error is reconducted via a duality argument to estimates for the backward and forward auxiliary problems. For the error in the dual variable, we use the following notation

$$\begin{aligned} \varepsilon &= w - w_{kh} \\ &= w - w_k + w_k - w_{kh} \\ &= \varepsilon_k + \varepsilon_h. \end{aligned}$$

When no confusion arises, we use the same notation for v, v_k, v_{kh} as well. The most challenging part consists in the L^∞ -norm estimate for the temporal error. In particular, the error arising from the time discretization is decomposed on each time interval I_n into an interpolation error and the error in the interior of the time interval, namely

$$\begin{aligned} e_k &= u - u_k \\ &= u(\cdot) - u(t_n) + u(t_n) - u_k(t_n), \end{aligned}$$

for $t_n \in I_n$.

The interpolation error, i.e., $u(\cdot) - u(t_n)$, is handled using the continuous auxiliary Problems 4.1.1 and 4.1.2 with corresponding energy-type estimates.

For the error inside the time interval, i.e., $u(t_n) - u_k(t_n)$, we introduce the following projection operator acting in time, see [84, Section 12].

Definition 4.1.10. Let $\pi_k: C(\bar{I} \setminus I_N, W) \rightarrow U_k(W)$ be the projection operator defined by

$$\pi_k w|_{I_n} = w(t_{n-1}),$$

where

(a) $W := H^{-1}(\Omega)$ for model problem (2.20);

(b) $W := H_0^1(\Omega)$ for model problem (2.27).

The extension of π_k to $U_k(W)$ is given by $\pi_k|_{U_k} = \mathbb{I}_{U_k}$, where \mathbb{I} is the identity operator.

A classical approximation property of this operator is given by

$$(4.20) \quad \|\varphi - \pi_k \varphi\|_I \leq C k_n^{r+1} \|\partial_t^{r+1} \varphi\|_{I_n},$$

see [84, Equation (12.10)], where the order r is the order of the discontinuous Galerkin approximation. Namely, $r = 0$ in our case.

This operator is used in the backward problems to express the error as

$$\begin{aligned} (4.21) \quad \varepsilon_k &= w - w_k \\ &= w - \pi_k w + \pi_k w - w_k \\ &= \eta_k + \xi_k, \end{aligned}$$

that is, a projection error η_k and a discretization error ξ_k . For the former, we exploit time-weighted estimate for the solution of Problem 4.1.1; the latter requires some more technical intermediates steps.

In the analysis of the spatial error e_h , and of the associated dual error ε_h , we use a similar approach. The solutions w_k and w_{kh} are not compared directly, we rather introduce the standard L^2 -projector in space

$$(4.22) \quad P_h: H_0^1(\Omega) \rightarrow V_h, \quad (P_h w_k, \varphi) = (w_k, \varphi), \quad \forall \varphi \in V_h,$$

and write

$$\begin{aligned} \varepsilon_h &= w_k - w_{kh} \\ &= w_k - P_h w_k + P_h w_k - w_{kh} \\ &= \eta_h + \xi_h. \end{aligned}$$

Then, we exploit the approximation properties of P_h , together with estimates for the solution w_k , to bound the projection error η_h . The second part, ξ_h , is purely a discretization error and the derivation requires several stability estimates for the forward auxiliary problems.

4.2 Stability analysis

We obtain stability results for the solutions of the auxiliary problems presented in the previous section. These results constitute the basis for the derivation of the estimates in negative norms in Sections 4.3 and 4.4.

4.2.1 Linear problem

In a first step, we obtain time-weighted estimates for the solution w of Problem 4.1.1 (b). This result corresponds to [56, Lemma 4.4].

Proposition 4.2.1. *Let $w \in L^2(I, L^2(\Omega)) \cap H^1(I, (\dot{H}^2)^*)$ be solution of Problem 4.1.1 (b). Then there holds*

$$(4.23a) \quad \int_I (T-t) \|\partial_t w(t)\|_{H^{-1}(\Omega)}^2 dt \leq C \|w_T\|_{H^{-1}(\Omega)}^2,$$

$$(4.23b) \quad \int_{I \setminus I_N} \|\partial_t w(t)\|_{H^{-1}(\Omega)} dt \leq C \left(\log \frac{T}{k} \right)^{\frac{1}{2}} \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. We start with the first relation testing (4.3) with $\varphi = (T-t)\Delta^{-1}\partial_t w$ to get

$$(4.24) \quad - \int_I (T-t) (\Delta^{-1} \partial_t w, \partial_t w) dt + \int_I (T-t) (\nabla \Delta^{-1} \partial_t w, \nabla w) dt = 0.$$

We apply integration by parts in space to both terms obtaining

$$(4.25) \quad \begin{aligned} - \int_I (T-t) (\Delta^{-1} \partial_t w, \partial_t w) dt &= - \int_I (T-t) (\Delta^{-1} \partial_t w, \Delta \Delta^{-1} \partial_t w) dt \\ &= \int_I (T-t) \|\nabla \Delta^{-1} \partial_t w\|^2 dt, \end{aligned}$$

and

$$(4.26) \quad \int_I (T-t) (\nabla \Delta^{-1} \partial_t w, \nabla w) = - \int_I (T-t) (\partial_t w, w) dt$$

For the latter, we observe

$$\int_I (T-t) (\partial_t w, w) dt = \frac{1}{2} \int_I \frac{d}{dt} ((T-t) \|w(t)\|^2) dt + \frac{1}{2} \|w\|_I^2$$

which, inserted back in (4.26), leads to

$$(4.27) \quad \begin{aligned} - \int_I (T-t) (\partial_t w, w) dt &= -\frac{1}{2} \|w\|_I^2 - \frac{1}{2} \int_I \frac{d}{dt} ((T-t) \|w(t)\|^2) dt \\ &= -\frac{1}{2} \|w\|_I^2 + \frac{T}{2} \|w(0)\|^2. \end{aligned}$$

Combining (4.25) with (4.27), we conclude from (4.24)

$$\int_I (T-t) \|\nabla \Delta^{-1} \partial_t w\|^2 dt + \frac{T}{2} \|w(0)\|^2 = \frac{1}{2} \|w\|_I^2 \leq C \|\nabla \Delta^{-1} w_T\|^2$$

where in the last step we used (3.7).

The last relation directly follows from (4.23a) thanks to the Cauchy-Schwarz inequality. In fact, recalling that $k \neq T$, there holds

$$\begin{aligned} \int_{I \setminus I_N} \|\partial_t w(t)\|_{H^{-1}(\Omega)} dt &\leq \left(\int_{I \setminus I_N} (T-t)^{-1} dt \right)^{\frac{1}{2}} \left(\int_{I \setminus I_N} (T-t) \|\partial_t w(t)\|_{H^{-1}(\Omega)}^2 dt \right)^{\frac{1}{2}} \\ &\leq C \left(\log \frac{T}{k} \right)^{\frac{1}{2}} \left(\int_I (T-t) \|\partial_t w(t)\|_{H^{-1}(\Omega)}^2 dt \right)^{\frac{1}{2}} \\ &\leq C \left(\log \frac{T}{k} \right)^{\frac{1}{2}} \|w_T\|_{H^{-1}}. \end{aligned}$$

□

We now extend the energy estimate (3.7) to the solution of the time-discrete Problem 4.1.4 (a) as well. This estimate has been derived in [56, Lemma 4.9].

Proposition 4.2.2. *Let $w_k \in U_k(H^{-1}(\Omega))$ be solution of (4.5). Then there holds*

$$(4.28) \quad \|w_k\|_I + \|w_{k,1}\|_{H^{-1}(\Omega)} \leq c \|w_T\|_{H^{-1}(\Omega)}$$

Proof. We observe that the bilinear form defined in (2.15) can be formulated as

$$\begin{aligned} B(\varphi_k, w_k) &= - \sum_{n=1}^N (\varphi_k, \partial_t w_k)_{I_n} + (\nabla \varphi_k, \nabla w_k)_I - \sum_{n=1}^{N-1} (\varphi_{k,n}, [w_k]_n) + (\varphi_{k,N}, w_{k,N}) \\ &= (\nabla \varphi_k, \nabla w_k)_I - \sum_{n=1}^{N-1} (\varphi_{k,n}, [w_k]_n) + (\varphi_{k,N}, w_{k,N}) \end{aligned}$$

by means of integration by parts in time.

Then, we rewrite (4.5) on any time interval I_n , $n = 1, \dots, N-1$, as

$$(4.29) \quad (\nabla \varphi_k, \nabla w_k)_{I_n} - (\varphi_{k,n}, [w_k]_n) = 0, \quad \forall \varphi_k \in \mathcal{P}_0(I_n, V_h),$$

observing that $w_{k,N} = w_T$.

In the expression above, we set as test function $\varphi_k = -\Delta^{-1} w_k$, obtaining

$$(4.30) \quad -(\nabla \Delta^{-1} w_k, \nabla w_k)_{I_n} + (\Delta^{-1} w_{k,n}, [w_k]_n) = 0.$$

Employing integration by parts in space to both terms, we obtain

$$(4.31) \quad -(\nabla \Delta^{-1} w_k, \nabla w_k)_{I_n} = (\Delta \Delta^{-1} w_k, w_k)_{I_n} = \|w_k\|_{I_n}^2.$$

and

$$\begin{aligned} (\Delta^{-1} w_{k,n}, [w_k]_n) &= (\Delta^{-1} w_{k,n}, [\Delta \Delta^{-1} w_k]_n) \\ &= -(\nabla \Delta^{-1} w_{k,n}, [\nabla \Delta^{-1} w_k]_n) \end{aligned}$$

For the latter, we exploit the relation

$$(4.32) \quad -(\varphi_{k,n}, [\varphi_k]_n) = \frac{1}{2} (-\|\varphi_{k,n+1}\|^2 + \|[\varphi_k]_n\|^2 + \|\varphi_{k,n}\|^2),$$

for all $\varphi_k \in U_k(V)$ to get, up to a constant,

$$(4.33) \quad -(\nabla \Delta^{-1} w_{k,n}, [\nabla \Delta^{-1} w_k]_n) \geq \|\nabla \Delta^{-1} w_{k,n}\|^2 - \|\nabla \Delta^{-1} w_{k,n+1}\|^2,$$

neglecting the positive jump-term in (4.32).

Inserting (4.31) and (4.33) in (4.30), it follows

$$\|w_k\|_{I_n}^2 + \|\nabla \Delta^{-1} w_{k,n}\|^2 - \|\nabla \Delta^{-1} w_{k,n+1}\|^2 \leq 0,$$

and the assertion is shown summing over $n = 1, \dots, N-1$. \square

Last results for this section is a stability results for the forward Problem 4.1.5 (a) already present in the literature, see [61, Theorem 4.6]. The proof will be given in Proposition 4.2.9 where this result will be extended to the semi-linear case.

Proposition 4.2.3. *Let v_k be solution of (4.7). Then there holds*

$$T\|\nabla \Delta v_{k,N}\|^2 + \|\Delta v_{k,N}\|^2 + \|\nabla \Delta v_k\|_I^2 + \sum_{n=2}^N \frac{t_{n-1}}{k_n} \|[\Delta v_k]_{n-1}\|^2 \leq C\|\Delta v_0\|^2.$$

\square

4.2.2 Extension to semi-linear problems

Firstly, similarly to Proposition 4.2.1, we obtain time-weighted estimates for the solution of Problem 4.1.6 (a) together with an energy estimate for it.

Proposition 4.2.4. *Let $w \in W(0, T)$ be solution of (4.11). Then there holds*

$$(4.34a) \quad \|\partial_t w\|_I + \|\nabla w\|_I + \max_{t \in I} \|w(t)\| \leq C\|w_T\|,$$

$$(4.34b) \quad \int_I (T-t) \|\partial_t w(t)\|^2 dt \leq C\|w_T\|^2,$$

$$(4.34c) \quad \int_{I \setminus I_N} \|\partial_t w(t)\| dt \leq C \left(\log \frac{T}{k} \right) \|w_T\|.$$

Proof. The first relation is already in the literature, see, e.g., [85, Lemma 7.10] and the reference therein.

We focus on (4.34b) and, through the choice $\varphi = -(T-t)\partial_t w$ in (4.11), we have

$$\begin{aligned} \int_I (T-t) \|\partial_t w\|^2 dt - \int_I (T-t) (\nabla \partial_t w, \nabla w) dt &= ((T-t)\partial_t w, \tilde{d}w)_I \\ &\leq \frac{1}{2} \left(\|(T-t)\partial_t w\|_I^2 + \|(T-t)\tilde{d}w\|_I^2 \right) \\ &\leq \frac{1}{2} \left(\|(T-t)\partial_t w\|_I^2 + c\|\tilde{d}\|_{\infty, \infty} \|w\|_I^2 \right), \end{aligned}$$

using Young's inequality and the boundedness of \tilde{d} .

Then, we observe that

$$\int_I (T-t) (\nabla \partial_t w, \nabla w) dt = \frac{1}{2} \int_I \frac{\partial}{\partial t} ((T-t) \|\nabla w(t)\|^2) dt + \frac{1}{2} \|\nabla w\|_I^2,$$

and conclude

$$\begin{aligned} \int_I (T-t) \|\partial_t w\|^2 dt + T \|\nabla w(0)\| &\leq \|\nabla w\|_I^2 + c \|\tilde{d}\|_{\infty, \infty} \|w\|_I^2 \\ &\leq C \|w_T\| \end{aligned}$$

thanks to (4.34a).

Once (4.34b) has been derived, the last stability estimate (4.34c) easily follows as in (4.23b) applying the Cauchy-Schwarz inequality, compare with the proof of Proposition 4.2.1 \square

Next result is concerned with the stability of the forward time-discrete auxiliary problem defined by (4.15). It will be used in Lemma 4.4.17 to obtain a $L^1(I, L^2(\Omega))$ -norm estimate for the solutions of Problem 4.1.7. This has to be compared with [61, Theorem 4.5] where the same it is obtained in the linear setting.

Proposition 4.2.5. *For the solution v_k of Problem 4.1.8 (a), there holds*

$$(4.35) \quad \|v_{k,N}\|^2 + T \|\nabla v_{k,N}\|^2 + \|\nabla v_k\|_I^2 + \sum_{n=1}^N t_n \|\Delta v_k\|_{I_n}^2 \leq C \|v_0\|^2.$$

Proof. On each time interval the equation (4.15) reads

$$(4.36) \quad (\nabla v_k, \nabla \varphi_k)_{I_n} + ([v_k]_{n-1}, \varphi_{k,n}) + (\hat{d}v_k, \varphi_k)_{I_n} = 0.$$

We test it with $\varphi_k = v_k$ and, using the relation

$$(4.37) \quad ([v_k]_{n-1}, v_{k,n}) = \frac{1}{2} (-\|v_{k,n-1}\|^2 + \|[v_k]_{n-1}\|^2 + \|v_{k,n}\|^2),$$

we get

$$(4.38) \quad 2\|\nabla v_k\|_{I_n}^2 + \|v_{k,n}\|^2 + 2\|\sqrt{\hat{d}}v_k\|_{I_n}^2 \leq \|v_{k,n-1}\|^2.$$

Summing for $n = 1, \dots, N$ we obtain the first part of the claim

$$(4.39) \quad 2\|\nabla v_k\|_I^2 + \|v_{k,N}\|^2 + 2\|\sqrt{\hat{d}}v_k\|_I^2 \leq \|v_{k,0}\|^2 = \|v_0\|^2.$$

To obtain the remaining estimate, we test (4.36) with $\varphi_k = -t_n \Delta v_k$ and get

$$-t_n (\nabla v_k, \nabla \Delta v_k)_{I_n} - t_n ([v_k]_{n-1}, \Delta v_{k,n}) - t_n (\hat{d}v_k, \Delta v_k)_{I_n} = 0.$$

Then we integrate by parts and use (4.37) to have

$$(4.40) \quad 2t_n \|\Delta v_k\|_{I_n}^2 + t_n \|\nabla v_{k,n}\|^2 \leq t_n \|\nabla v_{k,n-1}\|^2 + 2t_n (\hat{d}v_k, \Delta v_k)_{I_n}.$$

We focus on the right-hand side. Observing $t_n = t_{n-1} + k_n$ and recalling $k_n \leq \tilde{k}^{-1} k_{n-1}$, the former term is written as

$$(4.41) \quad \begin{aligned} t_n \|\nabla v_{k,n-1}\|^2 &= (t_{n-1} + k_n) \|\nabla v_{k,n-1}\|^2 \\ &\leq t_{n-1} \|\nabla v_{k,n-1}\|^2 + \tilde{k}^{-1} k_{n-1} \|\nabla v_{k,n-1}\|^2, \end{aligned}$$

while the latter, via Young's inequality and the boundedness of \hat{d} , takes form

$$(4.42) \quad 2t_n(\hat{d}v_k, \Delta v_k)_{I_n} \leq Ct_n\|v_k\|_{I_n}^2 + t_n\|\Delta v_k\|_{I_n}^2$$

Combing (4.41) and (4.42) in (4.40), summing over $n = 2, \dots, N$, and noting $t_1 = k_1$, we have

$$(4.43) \quad \begin{aligned} \sum_{n=2}^N t_n\|\Delta v_k\|_{I_n}^2 + T\|\nabla v_{k,N}\|^2 &\leq t_1\|\nabla v_{k,1}\|^2 + \tilde{k}^{-1} \sum_{n=2}^N k_{n-1}\|\nabla v_{k,n-1}\|^2 \\ &\quad + C \sum_{n=2}^N t_n\|v_k\|_{I_n}^2 \\ &= k_1\|\nabla v_{k,1}\|^2 + \tilde{k}^{-1} \sum_{n=1}^{N-1} \|\nabla v_k\|_{I_n}^2 \\ &\quad + C \sum_{n=2}^N t_n\|v_k\|_{I_n}^2 \\ &\leq C \left(\|\nabla v_k\|_I^2 + \sum_{n=2}^N t_n\|v_k\|_{I_n}^2 \right) \end{aligned}$$

The first term in the right-hand side displays the sought bound thanks to (4.39). On the other hand, we need a bound for the other term in the right-hand side as well as an estimate for $t_1\|\Delta v_k\|_{I_1}^2$ in the left-hand side. We start with the former.

(i) Testing (4.36) with $\varphi_k = t_n v_k$, using (4.37) and $k_n \leq \tilde{k}^{-1} k_{n-1}$, we have

$$\begin{aligned} 2t_n\|\nabla v_k\|_{I_n}^2 + 2t_n\|\sqrt{\hat{d}}v_k\|_{I_n}^2 + t_n\|v_{k,n}\|^2 &\leq t_{n-1}\|v_{k,n-1}\|^2 \\ &= (t_{n-1} + k_n)\|v_{k,n-1}\|^2 \\ &\leq t_{n-1}\|v_{k,n-1}\|^2 + \tilde{k}^{-1} k_{n-1}\|v_{k,n-1}\|^2. \end{aligned}$$

Summing for $n = 2, \dots, N$ yields

$$(4.44) \quad \begin{aligned} 2 \sum_{n=2}^N t_n\|\nabla v_k\|_{I_n}^2 + \sum_{n=2}^N t_n\|\sqrt{\hat{d}}v_k\|_{I_n}^2 + T\|v_{k,N}\|^2 \\ \leq t_1\|v_{k,0}\|^2 + \tilde{k}^{-1} \sum_{n=2}^N k_{n-1}\|v_{k,n-1}\|^2 \\ = k_1\|v_{k,0}\|^2 + \tilde{k}^{-1} \sum_{n=1}^{N-1} \|v_k\|_{I_n}^2 \\ \leq C\|v_k\|_I^2 \leq C\|\nabla v_k\|_I^2 \\ \leq C\|v_0\|^2 \end{aligned}$$

using (4.39) in the last step.

Then the claim follows observing that

$$(4.45) \quad \sum_{n=2}^N t_n\|v_k\|_{I_n}^2 \leq \sum_{n=2}^N t_n\|\nabla v_k\|_{I_n}^2 \leq C\|v_0\|^2.$$

- (ii) To bound $t_1 \|\Delta v_k\|_{I_1}^2$ we test (4.36) with $\varphi_k = -\Delta v_k$ restricting our attention on the first time interval. This gives, after integration by parts and thanks to \hat{d} being bounded,

$$\begin{aligned} \|\Delta v_k\|_{I_1}^2 &= ([v_k]_0, \Delta v_{k,1}) + (\hat{d}v_k, \Delta v_k)_{I_1} \\ &= (v_{k,1} - v_0, \Delta v_{k,1}) + (\hat{d}v_k, \Delta v_k)_{I_1} \\ &\leq \|v_{k,1} - v_0\| \|\Delta v_{k,1}\| + C \|v_k\|_{I_1} \|\Delta v_k\|_{I_1} \\ &= \|v_{k,1} - v_0\| \frac{1}{\sqrt{k_1}} \|\Delta v_k\|_{I_1} + C \|v_k\|_{I_1} \|\Delta v_k\|_{I_1}, \end{aligned}$$

which implies

$$\begin{aligned} k_1 \|\Delta v_{k,1}\| &\leq \|v_{k,1} - v_0\| + C \sqrt{k_1} \|v_k\|_{I_1} \\ &= \|v_{k,1} - v_0\| + C k_1 \|v_{k,1}\|. \end{aligned}$$

Then, using the inequality above, $t_1 = k_1$ and Minkowski's inequality, we have

$$\begin{aligned} t_1 \|\Delta v_k\|_{I_1}^2 &= t_1 k_1 \|\Delta v_{k,1}\|^2 = k_1^2 \|\Delta v_{k,1}\|^2 \\ (4.46) \quad &\leq C \left(\|v_{k,1}\|^2 + \|v_0\|^2 + k_1^2 \|v_{k,1}\|^2 \right) \\ &= C \left(\|v_{k,1}\|^2 + \|v_0\|^2 + t_1 \|v_k\|_{I_1}^2 \right) \end{aligned}$$

Then, the first term in the right-hand side is bounded by (4.38) restricted to $n = 1$, while for the last we observe that, summing for $n = 1, \dots, N$ in (4.44) gives

$$2 \sum_{n=1}^N t_n \|\nabla v_k\|_{I_n}^2 \leq \tilde{k}^{-1} \sum_{n=1}^N k_{n-1} \|v_{k,n-1}\|^2 \leq C \|v_k\|_I^2 \leq C \|\nabla v_k\|_I^2,$$

and, in particular,

$$t_1 \|v_k\|_{I_1}^2 \leq C t_1 \|\nabla v_k\|_{I_1}^2 \leq C \|\nabla v_k\|_I^2 \leq C \|v_0\|^2$$

using (4.39) in the last inequality.

Back in (4.46), this gives

$$(4.47) \quad t_1 \|\Delta v_k\|_{I_1}^2 \leq C \|v_0\|$$

Combining (4.45) and (4.47) in (4.43), we obtain the claim

$$(4.48) \quad \sum_{n=1}^N t_n \|\Delta v_k\|_{I_n}^2 + T \|\nabla v_{k,N}\|^2 \leq C \|v_0\|.$$

□

Remark 4.2.6. The part (i) of the proof above can be obtained directly through the Poincaré inequality. We have not use it in the first place in order to derive inequality (4.44), which has been later used in the proof.

Next, we need accessory results to handle the derivatives of the term $d(\cdot, \cdot, u)$. It is here that we need the semi-linear term to be four times differentiable, see Remark 2.3.1. In the following, we shorten the notation to $d(u)$.

Lemma 4.2.7. *For $u \in L^2(I, H^2(\Omega)) \cap L^\infty(I, H_0^1(\Omega))$ there holds*

$$(4.49) \quad \|\Delta \partial_u^i d(u)\|_I + \|\nabla \partial_u^i d(u)\|_{L^4(I \times \Omega)} \leq C, \quad i = 1, 2.$$

where the constant C depends on $\|u\|_{H^2(\Omega)}$.

Proof. With no loss of generality we restrict our attention on the spatial variable x rather than on the whole spatial domain, and we show the claim for $i = 2$. Referring to [85, Section 4.3.2] for the notion of derivative for a Nemytskii operator, we have

$$(4.50) \quad \partial_x^2 \partial_u^2 d(u) = \partial_u^3 d(u) \partial_x^2 u + \partial_u^4 d(u) \partial_x u \partial_x u + \partial_x d(u) + \partial_u d(u) \partial_x u.$$

By Assumption 2.3.1 (i) the Nemytskii operator d is fourth continuously differentiable with respect to u . Further, the assumed spatial regularity of u and the embedding $L^\infty(I, H_0^1(\Omega)) \hookrightarrow L^4(I \times \Omega)$ implies $\partial_x u \in L^4(I \times \Omega)$. Therefore, we can regard the right-hand side of (4.50) as an element of $L^2(I, L^2(\Omega))$, from which we infer

$$\|\Delta \partial_u^2 d(u)\|_I \leq C \|\Delta u\|_I + C \|\nabla u\|_{L^4(I \times \Omega)}^2.$$

For the term left, we have

$$\partial_x \partial^2 d(u) = \partial_u^3 d(u) \partial_x u,$$

and therefore the claim follows with same arguments as before. \square

Lemma 4.2.8. *For the solution v_k of Problem 4.1.8(a), there holds*

$$(4.51) \quad \|\Delta(\hat{d}v_k)\|_{I_n} \leq C(\|v_k\|_{I_n} + \|\nabla v_k\|_{L^4(I \times \Omega)} + \|\Delta v_k\|_{I_n})$$

for any $n = 1, \dots, N$.

Proof. We observe that by definition \hat{d} is the difference quotient of the semi-linear term d . We have

$$\hat{d} = \frac{d(u) - d(u_k)}{u - u_k} = \partial_u d(u) + \frac{1}{2} \int_0^1 \partial_u^2 d(u + s(u_k - u))(u - u_k) ds.$$

Then, with $v(s) = u + s(u_k - u)$, it follows

$$(4.52) \quad \Delta(\hat{d}v_k) = \Delta(\partial_u d(u)v_k) + \frac{1}{2} \int_0^1 \Delta(\partial_u^2 d(v(s))(u - u_k)v_k) ds.$$

For the former term in the right-hand side of (4.52) we have

$$\Delta(\partial_u d(u)v_k) = \partial_u d(u) \Delta v_k + 2 \nabla \partial_u d(u) \nabla v_k + v_k \Delta \partial_u d(u),$$

from which

$$(4.53) \quad \begin{aligned} \|\Delta(\partial_u d(u)v_k)\|_{I_n} &\leq \|\partial_u d(u)\|_{\infty, \infty} \|\Delta v_k\|_{I_n} + \|\Delta \partial_u d(u)\|_{I_n} \|v_k\|_{I_n} \\ &\quad + 2 \|\nabla \partial_u d(u)\|_{L^4(I_n \times \Omega)} \|\nabla v_k\|_{L^4(I_n \times \Omega)} \\ &\leq C(\|v_k\|_{I_n} + \|\nabla v_k\|_{L^4(I \times \Omega)} + \|\Delta v_k\|_{I_n}), \end{aligned}$$

where in the last step, we used Assumption 2.3.1 to bound $\|\partial_u d(u)\|_{\infty, \infty}$ and Lemma 4.2.7 for the rest.

For the remaining term in (4.52), we have

$$\begin{aligned} \int_0^1 \Delta(\partial_u^2 d(v(s))(u - u_k)v_k) ds &= \int_0^1 \partial_u^2 d(v(s)) \Delta((u - u_k)v_k) ds \\ &\quad + 2 \int_0^1 \nabla \partial_u^2 d(v(s)) \nabla((u - u_k)v_k) ds \\ &\quad + \int_0^1 \Delta(\partial_u^2 d(v(s)))(u - u_k)v_k ds. \end{aligned}$$

Applying Lemma 4.2.7 to the equality above, we obtain

$$\begin{aligned} (4.54) \quad \left\| \int_0^1 \Delta(\partial_u^2 d(v(s))(u - u_k)v_k) ds \right\|_{I_n} &\leq C \|u - u_k\|_{\infty, \infty} \|v_k\|_{I_n} \\ &\quad + C \|\nabla((u - u_k))\|_{L^4(I_n \times \Omega)} \|\nabla v_k\|_{L^4(I_n \times \Omega)} \\ &\quad + C \|\Delta((u - u_k))\|_{I_n} \|\Delta v_k\|_{I_n} \\ &\leq C (\|v_k\|_{I_n} + \|\nabla v_k\|_{L^4(I \times \Omega)} + \|\Delta v_k\|_{I_n}), \end{aligned}$$

using in the last step the stability of u and u_k solutions of (2.28) and (2.29) from Proposition 3.2.2 and 3.2.3, respectively.

Then the claim follows combining (4.53) with (4.54). \square

Last result of this section is a stability estimate similar to the one of Proposition 4.2.5 but this time with respect to the $H^2(\Omega)$ -norm of the initial data. This will be needed in Lemmas 4.4.14 and 4.4.15. The following stability estimate has been derived in the linear setting in [61, Theorem 4.6]. For the extension to the semi-linear setting, the previous Lemma 4.2.8 is required to deal with the presence of the semi-linear term.

Proposition 4.2.9. *For the solution v_k of Problem 4.1.8(a), it holds*

$$\begin{aligned} (4.55) \quad T \|\nabla \Delta v_{k,N}\|^2 + \|\Delta v_{k,N}\|^2 + \sum_{n=2}^N \frac{t_{n-1}}{k_n} \|[\Delta v_k]_{n-1}\|^2 + \sum_{n=2}^N t_{n-1} \|[\Delta v_k]_{n-1}\|^2 \\ \leq C \|\Delta v_0\|^2 \end{aligned}$$

Proof. In a first step, after rewriting (4.15) on each time interval I_n , we test it with $\varphi_k = \Delta^2 v_k$ and, after integration by parts in the first term, we have

$$(4.56) \quad \|\nabla \Delta v_k\|_{I_n}^2 + ([\Delta v_k]_{n-1}, \Delta v_{k,n}) + (\hat{d}v_k, \Delta^2 v_k)_{I_n} = 0.$$

Then, thanks to (4.37), we have

$$(4.57) \quad 2 \|\nabla \Delta v_k\|_{I_n}^2 + \|\Delta v_{k,n}\|^2 \leq \|\Delta v_{k,n-1}\|^2 - 2(\Delta(\hat{d}v_k), \Delta v_k)_{I_n}.$$

For the last term in the right hand side, we use Lemma 4.2.8 to conclude

$$\begin{aligned} -2(\Delta(\hat{d}v_k), \Delta v_k)_{I_n} &\leq \|\Delta(\hat{d}v_k)\|_{I_n} \|\Delta v_k\|_{I_n} \\ &\leq C \|\Delta v_k\|_{I_n}^2, \end{aligned}$$

4.2. Stability analysis

using the fact that the $\|\cdot\|_{I_n}$ and the $\|\cdot\|_{L^4(I \times \Omega)}$ norms are bounded by the $\|\cdot\|_{L^2(I, H^2(\Omega))}$ norm.

Inserting the inequality above in (4.57) and summing for $n = 1, \dots, N$, we get

$$(4.58) \quad 2\|\nabla \Delta v_k\|_I^2 + \|\Delta v_{k,N}\|^2 \leq \|\Delta v_0\|^2 + C\|\Delta v_k\|_I^2 \leq C\|\Delta v_0\|^2.$$

To estimate the jump term, we test (4.15) with $\varphi_k = \frac{t_{n-1}}{k_n}[\Delta^2 v_k]_{n-1}$ and obtain, after integration by parts in the first term,

$$\frac{t_{n-1}}{k_n}(\nabla \Delta v_k, [\nabla \Delta v_k]_{n-1})_{I_n} + \frac{t_{n-1}}{k_n}\|[\Delta v_k]_{n-1}\|^2 + (\hat{d}v_k, [\Delta^2 v_k]_{n-1})_{I_n} = 0.$$

Then, observing that

$$\frac{t_{n-1}}{k_n}(\nabla \Delta v_k, [\nabla \Delta v_k]_{n-1})_{I_n} = t_{n-1}(\nabla \Delta v_k, [\nabla \Delta v_k]_{n-1}),$$

we apply relation (4.37) to this term, and get

$$(4.59) \quad \begin{aligned} t_n\|\nabla \Delta v_{k,n}\|^2 + 2\frac{t_{n-1}}{k_n}\|[\Delta v_k]_{n-1}\|^2 &\leq t_{n-1}\|\nabla \Delta v_{k,n-1}\|^2 + k_n\|\nabla \Delta v_{k,n}\|^2 \\ &\quad - 2\frac{t_{n-1}}{k_n}(\Delta(\hat{d}v_k), [\Delta v_k]_{n-1})_{I_n}, \end{aligned}$$

using that $t_{n-1} = t_n - k_n$. We focus on the last term in the right hand-side. Young's inequality yields

$$\begin{aligned} -2\frac{t_{n-1}}{k_n}(\Delta(\hat{d}v_k), [\Delta v_k]_{n-1})_{I_n} &\leq t_{n-1}\|\Delta(\hat{d}v_k)\|_{I_n}^2 + \frac{t_{n-1}}{k_n^2}\|[\Delta v_k]_{n-1}\|_{I_n}^2 \\ &\leq T\|\Delta(\hat{d}v_k)\|_{I_n}^2 + \frac{t_{n-1}}{k_n}\|[\Delta v_k]_{n-1}\|^2 \end{aligned}$$

Then, arguing as in the first part of the proof, we have

$$(4.60) \quad \|\Delta(\hat{d}v_k)\|_{I_n}^2 \leq C\|\Delta v_k\|_{I_n}^2,$$

and ultimately

$$-2\frac{t_{n-1}}{k_n}(\Delta(\hat{d}v_k), [\Delta v_k]_{n-1})_{I_n} \leq C\|\Delta v_k\|_{I_n}^2 + \frac{t_{n-1}}{k_n}\|[\Delta v_k]_{n-1}\|^2.$$

Using the relation above in (4.59) and summing for $n = 2, \dots, N$ we obtain

$$(4.61) \quad \begin{aligned} T\|\nabla \Delta v_{k,N}\|^2 + \sum_{n=2}^N \frac{t_{n-1}}{k_n}\|[\Delta v_k]_{n-1}\|^2 &\leq t_1\|\nabla \Delta v_{k,1}\|^2 \\ &\quad + \sum_{n=2}^N \|\nabla \Delta v_k\|_{I_n}^2 + C \sum_{n=2}^N \|\Delta v_k\|_{I_n}^2 \\ &\leq \|\nabla \Delta v_k\|_I^2 + C\|\Delta v_k\|_I^2, \end{aligned}$$

using in the last step $t_1 = k_1$.

Then, combining the relation above with (4.58), we obtain the claim. \square

4.3 Error analysis for linear equations

Based on the stability analysis of Section 4.2, we derive $L^\infty(I, H_0^1(\Omega))$ -norm estimates for the solutions of the continuous state equation (2.22), and the semi-discrete (2.23) and discrete (2.24) one.

4.3.1 Temporal error

As we will see at the end of this section, the error e_k in the primal variable depends on the error in the dual variable ε_k in the $L^1(I, H^{-1}(\Omega))$ and $H^{-3}(\Omega)$ -norms. We state these estimates in the following theorem whose proof will be given subsequently with a series of lemmas. This result corresponds to [56, Lemma 4.7] of the thesis author.

Theorem 4.3.1. *Let w and w_k be solutions of Problems (4.1.1)(b) and (4.1.4)(a), respectively. Then, the corresponding error satisfies*

$$(4.62) \quad \|w - w_k\|_{L^1(I, H^{-1}(\Omega))} + \|w(0) - w_{k,1}\|_{H^{-3}(\Omega)} \leq Ck \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. The strategy is to compare w with the projection operator π_k when applied to w itself. Observing that $\pi_k w(0) = w(0)$, we split the error into

$$\begin{aligned} & \|w - w_k\|_{L^1(I, H^{-1}(\Omega))} + \|w(0) - w_{k,1}\|_{H^{-3}(\Omega)} \\ & \leq \|w - \pi_k w\|_{L^1(I, H^{-1}(\Omega))} + \|\pi_k w - w_k\|_{L^1(I, H^{-1}(\Omega))} + \|\pi_k w(0) - w_{k,1}\|_{H^{-3}(\Omega)}. \end{aligned}$$

Then, the claim follows combining Lemmas 4.3.3 and 4.3.5. \square

First lemma involves the continuous problems 4.1.1 and 4.1.2; it will be used to bound the error at the temporal nodal point. This result has been derived in [56, Lemma 4.3] by the author of this thesis.

Lemma 4.3.2. *For the error between w and \hat{w} solutions of Problems (4.1.1)(b) and (4.1.2)(b), respectively, there holds*

$$(4.63) \quad \|w - \hat{w}\|_{L^1(\hat{I}, H^{-1}(\Omega))} + \|w(0) - \hat{w}(0)\|_{H^{-3}(\Omega)} \leq Ck \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. To derive the equation for the error $\hat{\varepsilon} := \hat{w} - w$, we subtract (4.3) from (4.4) and integrate on \hat{I} only, obtaining

$$-(\varphi, \partial_t \hat{\varepsilon})_{\hat{I}} + (\nabla \varphi, \nabla \hat{\varepsilon})_{\hat{I}} = 0$$

for any $\varphi \in L^2(\hat{I}, \dot{H}^2(\Omega)) \cap H^1(\hat{I}, L^2(\Omega))$.

Integration by parts in the second term gives

$$(4.64) \quad -(\varphi, \partial_t \hat{\varepsilon})_{\hat{I}} - (\Delta \varphi, \hat{\varepsilon})_{\hat{I}} = 0.$$

We estimate separately the two terms in the left-hand side of (4.63), starting with $\|\hat{\varepsilon}(0)\|_{H^{-3}(\Omega)}$.

(i) Testing (4.64) with $\varphi = -\Delta^{-3}\hat{\varepsilon}$, we have

$$(4.65) \quad (\Delta^{-3}\hat{\varepsilon}, \partial_t \hat{\varepsilon})_{\hat{I}} + (\Delta^{-2}\hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} = 0.$$

For the term containing the time derivative, we observe that

$$(4.66) \quad \int_{\hat{I}} \partial_t (\Delta^{-3}\hat{\varepsilon}, \hat{\varepsilon}) dt = (\partial_t (\Delta^{-3}\hat{\varepsilon}), \hat{\varepsilon})_{\hat{I}} + (\Delta^{-3}\hat{\varepsilon}, \partial_t \hat{\varepsilon})_{\hat{I}},$$

implying

$$(\Delta^{-3}\hat{\varepsilon}, \partial_t \hat{\varepsilon})_{\hat{I}} = (\Delta^{-3}\hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})) - (\Delta^{-3}\hat{\varepsilon}(0), \hat{\varepsilon}(0)) - (\partial_t (\Delta^{-3}\hat{\varepsilon}), \hat{\varepsilon})_{\hat{I}}.$$

Substituting in (4.65), we get

$$(4.67) \quad \begin{aligned} & -(\Delta^{-3}\hat{\varepsilon}(0), \hat{\varepsilon}(0)) - (\partial_t (\Delta^{-3}\hat{\varepsilon}), \hat{\varepsilon})_{\hat{I}} + (\Delta^{-2}\hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} \\ & = -(\Delta^{-3}\hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})). \end{aligned}$$

We consider each term in the last equation separately. The self-adjointness of Δ^{-1} and integration by parts in space, yield

$$(4.68) \quad \begin{aligned} -(\Delta^{-3}\hat{\varepsilon}(0), \hat{\varepsilon}(0)) & = -(\Delta^{-2}\hat{\varepsilon}(0), \Delta\Delta^{-2}\hat{\varepsilon}(0)) \\ & = (\nabla\Delta^{-2}\hat{\varepsilon}(0), \nabla\Delta^{-2}\hat{\varepsilon}(0)) \\ & = \|\nabla\Delta^{-2}\hat{\varepsilon}(0)\|^2. \end{aligned}$$

For the second term in (4.67), we use the relation $\partial_t \hat{\varepsilon} = -\Delta\hat{\varepsilon}$ and again Δ^{-1} being self-adjoint, to write

$$(4.69) \quad -(\partial_t (\Delta^{-3}\hat{\varepsilon}), \hat{\varepsilon})_{\hat{I}} = (\Delta^{-2}\hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} = (\Delta^{-1}\hat{\varepsilon}, \Delta^{-1}\hat{\varepsilon})_{\hat{I}} = \|\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2,$$

while for the third term we easily obtain

$$(4.70) \quad (\Delta^{-2}\hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} = (\Delta^{-1}\hat{\varepsilon}, \Delta^{-1}\hat{\varepsilon})_{\hat{I}} = \|\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2.$$

To estimate the right-hand side of (4.67), we firstly use the same argument as in (4.68) to reformulate it as

$$(4.71) \quad \begin{aligned} -(\Delta^{-3}\hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})) & = -(\Delta^{-2}\hat{\varepsilon}(\hat{t}), \Delta^{-1}\hat{\varepsilon}(\hat{t})) \\ & = -(\Delta^{-2}\hat{\varepsilon}(\hat{t}), \Delta\Delta^{-2}\hat{\varepsilon}(\hat{t})) \\ & = \|\nabla\Delta^{-2}\hat{\varepsilon}(\hat{t})\|^2, \\ & = \int_{\Omega} \left(\int_{\hat{t}}^T \nabla\Delta^{-2}\partial_t w(t) dt \right)^2 dx \end{aligned}$$

where we have used that

$$\hat{\varepsilon}(\hat{t}) = w_T - w(\hat{t}).$$

Then, with the help of the Cauchy-Schwarz inequality, the Fubini-Tonelli theorem, the relation $\partial_t = -\Delta^{-1}$, and using that Δ^{-1} is self-adjoint, we

have

$$\begin{aligned}
 (4.72) \quad \int_{\Omega} \left(\int_{\hat{t}}^T \nabla \Delta^{-2} \partial_t w(t) dt \right)^2 dx &= \int_{\Omega} \left(\int_{\hat{t}}^T -\nabla \Delta^{-1} w(t) dt \right)^2 dx \\
 &= \int_{\Omega} \left(\int_{\hat{t}}^T 1(-\nabla \Delta^{-1} w(t)) dt \right)^2 dx \\
 &\leq \int_{\Omega} \left(\sqrt{\int_{\hat{t}}^T dt} \sqrt{\int_{\hat{t}}^T |-\nabla \Delta^{-1} w(t)|^2 dt} \right)^2 dx \\
 &\leq k \int_{\Omega} \left(\int_{\hat{t}}^T |\nabla \Delta^{-1} w(t)|^2 dt \right)^2 dx \\
 &= k \int_{\hat{t}}^T \int_{\Omega} |\nabla \Delta^{-1} w(t)|^2 dx dt.
 \end{aligned}$$

Finally, merging (4.71) with (4.72) and using (3.7), we conclude for the right-hand side of (4.67)

$$\begin{aligned}
 (4.73) \quad -(\Delta^{-3} \hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})) &= \|\nabla \Delta^{-2} \hat{\varepsilon}(\hat{t})\|^2 \\
 &\leq k \int_{\hat{t}}^T \int_{\Omega} |\nabla \Delta^{-1} w(t)|^2 dx dt \\
 &\leq Ck^2 \|\nabla \Delta^{-1} w_T\|^2.
 \end{aligned}$$

Combining (4.67) with the relations (4.68), (4.69), (4.70) and the estimate of the right-hand side (4.73), we conclude

$$(4.74) \quad \|\nabla \Delta^{-2} \hat{\varepsilon}(0)\|^2 + 2\|\Delta^{-1} \hat{\varepsilon}\|_{\hat{I}}^2 \leq Ck^2 \|\nabla \Delta^{-1} w_T\|^2.$$

- (ii) For the term $\|\hat{\varepsilon}\|_{L^1(I, H^{-1}(\Omega))}$, we introduce $\tau(t) = \max(\hat{t} - t, k)$ for $t \in \hat{I}$ and observe that, using Cauchy-Schwarz inequality, there holds

$$\begin{aligned}
 (4.75) \quad \|\hat{\varepsilon}\|_{L^1(\hat{I}, H^{-1}(\Omega))}^2 &\leq \|\sqrt{\tau}^{-1}\|_{L^2(\hat{I})}^2 \|\sqrt{\tau} \hat{\varepsilon}\|_{L^2(\hat{I}, H^{-1}(\Omega))}^2 \\
 &\leq C \left(\log \frac{T}{k} + 1 \right) \|\sqrt{\tau} \hat{\varepsilon}\|_{L^2(\hat{I}, H^{-1}(\Omega))}^2.
 \end{aligned}$$

Then, if we are able to show the relation

$$(4.76) \quad \|\sqrt{\tau} \hat{\varepsilon}\|_{L^2(\hat{I}, H^{-1}(\Omega))}^2 \leq Ck^2 \|w_T\|_{H^{-1}(\Omega)}^2,$$

we would obtain from (4.75) the claim

$$\|\hat{\varepsilon}\|_{L^1(\hat{I}, H^{-1}(\Omega))}^2 \leq Ck^2 \left(\log \frac{T}{k} + 1 \right) \|w_T\|_{H^{-1}(\Omega)}^2.$$

Therefore, we focus in the derivation of (4.76).

Testing (4.64) with $\varphi = \tau \Delta^{-2} \hat{\varepsilon}$, it follows

$$(4.77) \quad -(\tau \Delta^{-2} \hat{\varepsilon}, \partial_t \hat{\varepsilon})_{\hat{I}} - (\tau \Delta^{-1} \hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} = 0.$$

For the first term on the left-hand side, we use the relation

$$\begin{aligned} -(\tau \Delta^{-2} \hat{\varepsilon}, \partial_t \hat{\varepsilon})_{\hat{I}} &= -\frac{1}{2} \int_{\hat{I}} \partial_t \left(\tau (\Delta^{-2} \hat{\varepsilon}(t), \hat{\varepsilon}(t)) \right) dt + \frac{1}{2} \int_{\hat{I}} \tau' (\Delta^{-2} \hat{\varepsilon}(t), \hat{\varepsilon}(t)) dt \\ &= -\frac{1}{2} \left((k \Delta^{-2} \hat{\varepsilon}(\hat{t}), \hat{\varepsilon}(\hat{t}) - (\hat{t} \Delta^{-2} \hat{\varepsilon}(0), \hat{\varepsilon}(0)) \right) \\ &\quad + \frac{1}{2} \int_{\hat{I}} \tau' (\Delta^{-2} \hat{\varepsilon}(t), \hat{\varepsilon}(t)) dt \end{aligned}$$

where τ' denotes the first derivative of τ with respect to t .

The second term in (4.77) is handled by

$$-(\tau \Delta^{-1} \hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} = -(\sqrt{\tau} \Delta^{-1} \hat{\varepsilon}, \sqrt{\tau} \Delta \Delta^{-1} \hat{\varepsilon}) = \|\sqrt{\tau} \nabla \Delta^{-1} \hat{\varepsilon}\|_{\hat{I}}^2.$$

Then, observing that $-\tau' \leq 1$ and $\hat{\varepsilon}(\hat{t}) = w_T - w(\hat{t})$, we obtain from (4.77) and the two before-mentioned relations

$$\begin{aligned} (4.78) \quad & \frac{\hat{t}}{2} \|\Delta^{-1} \hat{\varepsilon}(0)\|^2 + \|\sqrt{\tau} \nabla \Delta^{-1} \hat{\varepsilon}\|_{\hat{I}}^2 \\ & \leq \frac{\|\Delta^{-1} \hat{\varepsilon}\|_{\hat{I}}^2}{2} + \frac{k}{2} (\Delta^{-2} \hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})). \end{aligned}$$

In the next step, we estimate the second term in the right-hand side of the previous expression. Integration by parts in space, the estimate in (4.73), and (3.7), lead to

$$\begin{aligned} k(\Delta^{-2} \hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})) &= k(\Delta^{-2} \hat{\varepsilon}(\hat{t}), \Delta \Delta^{-1} w_T - w(\hat{t})) \\ &= -k(\nabla \Delta^{-2} \hat{\varepsilon}(\hat{t}), \nabla \Delta^{-1} (w_T - w(\hat{t}))) \\ &\leq k \|\nabla \Delta^{-2} \hat{\varepsilon}(\hat{t})\| \|\nabla \Delta^{-1} (w_T - w(\hat{t}))\| \\ &\leq C k^2 \|\nabla \Delta^{-1} w_T\|^2. \end{aligned}$$

Then, from (4.78) and thanks to (4.74) we conclude

$$\hat{t} \|\Delta^{-1} \hat{\varepsilon}(0)\|^2 + 2 \|\sqrt{\tau} \nabla \Delta^{-1} \hat{\varepsilon}\|_{\hat{I}}^2 \leq C k^2 \|\nabla \Delta^{-1} w_T\|^2.$$

This establishes (4.76) as required. □

We move our attention to the projection error

$$\eta_k = w - \pi_k w,$$

taking advantage from the approximation property of π_k and the time-weighted estimates from Proposition 4.2.1. The next three results correspond to [56, Lemma 4.7] of the author of this thesis.

Lemma 4.3.3. *For the error between the solution w of Problem 4.1.1(b) and its projection in time there holds*

$$(4.79) \quad \|w - \pi_k w\|_{L^1(I, H^{-1}(\Omega))} \leq c k \left(\log \frac{T}{k} \right)^{\frac{1}{2}} \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. In view of the time-weighted estimate (4.23b), we split the projection error

$$\|w - \pi_k w\|_{L^1(I, H^{-1}(\Omega))} = \int_{I \setminus I_N} \|w - \pi_k w\|_{H^{-1}(\Omega)} dt + \int_{I_N} \|w - \pi_k w\|_{H^{-1}(\Omega)} dt.$$

For the latter, we clearly have

$$(4.80) \quad \begin{aligned} \int_{I_N} \|w - \pi_k w\|_{H^{-1}(\Omega)} dt &\leq ck \max_{t \in I_N} \|w(t)\|_{H^{-1}(\Omega)} \\ &\leq ck \|w_T\|_{H^{-1}(\Omega)}, \end{aligned}$$

using (3.7) in the last inequality.

For the former, we observe that the operator Δ^{-1} acts in space only and thus, being independent of t , interchanges with π_k . This yields

$$\begin{aligned} \int_{I \setminus I_N} \|w - \pi_k w\|_{H^{-1}(\Omega)} dt &= \int_{I \setminus I_N} \|\nabla \Delta^{-1}(w - \pi_k w)\| dt, \\ &= \int_{I \setminus I_N} \|\nabla(\Delta^{-1}w - \pi_k \Delta^{-1}w)\| dt. \end{aligned}$$

Then, by transformation to the reference-time element, we have

$$(4.81) \quad \int_{I \setminus I_N} \|w - \pi_k w\|_{H^{-1}(\Omega)} dt \leq ck \int_{I \setminus I_N} \|\partial_t w(t)\|_{H^{-1}(\Omega)} dt,$$

see also [84, Equation (12.10)].

Then, the claim follows using (4.23b) in (4.81) and combining it with (4.80). \square

We write now the expression for the discretization error

$$\xi_k = \pi_k w - w_k,$$

and bounded it afterward.

Lemma 4.3.4. *The discretization error ξ_k satisfies the equation*

$$(4.82) \quad (\nabla \varphi_k, \nabla \xi_k)_{I_n} - (\varphi_{k,n}, [\xi_k]_n) = \int_{I_n} (t_n - t)(\Delta \varphi_k, \partial_t w(t)) dt,$$

for any $i = 1, \dots, N$ and $\varphi_k \in \mathcal{P}_0(I_n, H^2(\Omega) \cap H_0^1(\Omega))$.

Proof. Exploiting the Galerkin orthogonality, the definition of π_k and the fact that (4.3) and (4.5) have the same initial value, we write

$$\begin{aligned} B(\varphi_k, \xi_k) &= B(\varphi_k, \pi_k w - w + w - w_k) = -B(\varphi_k, w - \pi_k w) \\ &= -(\nabla \varphi_k, \nabla(w - \pi_k w))_I - \sum_{n=2}^N ([\varphi_k]_{n-1}, w(t_{n-1}) - \pi_k w)_{I_n} \\ &\quad - (\varphi_{k,1}, w_T - w_T) \\ &= (\Delta \varphi_k, w - \pi_k w)_I, \end{aligned}$$

using integration by parts in the last step.

We now expand the expression above on each time interval, and, recalling that the size of I_n is $k_n = t_n - t_{n-1}$, we obtain

$$\begin{aligned} B(\varphi_k, \xi_k) &= \sum_{n=1}^N \left(\int_{I_n} (\Delta \varphi_{k,n}, w(t)) dt - k_n (\Delta \varphi_{k,n}, w(t_{n-1})) \right) \\ &= \sum_{n=1}^N \left(\int_{I_n} (t_n - t) (\Delta \varphi_{k,n}, \partial_t w(t)) dt \right). \end{aligned}$$

In conclusion, combining the expressions above and integrating by parts in time $B(\varphi_k, \xi_k)$, we obtain the claim

$$(\nabla \varphi_k, \nabla \xi_k)_{I_n} - (\varphi_{k,n}, [\xi_k]_n) = \int_{I_n} (t_n - t) (\Delta \varphi_k, \partial_t w(t)) dt,$$

on each time interval I_n , $n = 1, \dots, N$ □

Lemma 4.3.5. *Let w and w_k be solutions of Problems 4.1.1(b) and 4.1.4(a), respectively. Then, for the discretization error there holds*

$$(4.83) \quad \|\pi_k w(0) - w_{k,1}\|_{H^{-3}(\Omega)} + \|\pi_k w - w_k\|_{L^1(I, H^{-1}(\Omega))} \leq Ck \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. We start the analysis estimating $\|\pi_k w(0) - w_{k,1}\|_{H^{-3}(\Omega)}$. Recalling the abbreviation $\xi_k = \pi_k w - w_k$, we set $\varphi_k = -\Delta^{-3} \xi_k$ in (4.82) and use $\partial_t w = -\Delta w$, obtaining

$$\begin{aligned} &-(\nabla \Delta^{-3} \xi_k, \nabla \xi_k)_{I_n} + (\Delta^{-3} \xi_{k,n}, [\xi_k]_n) \\ (4.84) \quad &= - \int_{I_n} (t_n - t) (\Delta^{-2} \xi_k, \partial_t w(t)) dt \\ &= \int_{I_n} (t_n - t) (\Delta^{-1} \xi_k, w(t)) dt. \end{aligned}$$

Focusing on the left-hand side, the former term is integrated by parts

$$(4.85) \quad -(\nabla \Delta^{-3} \xi_k, \nabla \xi_k)_{I_n} = (\Delta^{-2} \xi_k, \xi_k)_{I_n} = \|\Delta^{-1} \xi_k\|_{I_n}^2.$$

For the latter, we use again integration by parts together with the equality (4.32), obtaining

$$\begin{aligned} &(\Delta^{-3} \xi_{k,n}, [\xi_k]_n) = -(\nabla \Delta^{-2} \xi_{k,n}, [\nabla \Delta^{-2} \xi_k]_n) \\ (4.86) \quad &\geq \frac{1}{2} (\|\nabla \Delta^{-2} \xi_{k,n}\|^2 - \|\nabla \Delta^{-2} \xi_{k,n+1}\|^2). \end{aligned}$$

Combining (4.85) with (4.86) in (4.84), we have

$$\begin{aligned} &\|\Delta^{-1} \xi_k\|_{I_n}^2 + \frac{1}{2} (\|\nabla \Delta^{-2} \xi_{k,n}\|^2 - \|\nabla \Delta^{-2} \xi_{k,n+1}\|^2) \\ (4.87) \quad &\leq \int_{I_n} (t_n - t) (\Delta^{-1} \xi_k, w(t)) dt \\ &\leq \int_{I_n} k (\Delta^{-1} \xi_k, w(t)) dt \\ &\leq \frac{1}{2} \|\Delta^{-1} \xi_k\|_{I_n}^2 + \frac{1}{2} \int_{I_n} k^2 \|w(t)\|^2 dt, \end{aligned}$$

using Young's inequality in the last step.

We now sum over $n = 1, \dots, N$, noting that $\xi_{k,N+1} = 0$, to conclude

$$(4.88) \quad \begin{aligned} \|\Delta^{-1}\xi_k\|_I^2 + \|\nabla\Delta^{-2}\xi_{k,1}\|^2 &\leq Ck^2\|w\|_I^2 \\ &\leq Ck^2\|\nabla\Delta^{-1}w_T\|^2, \end{aligned}$$

where in the last step we have used the stability estimate (3.7). This concludes the first part.

For the second part of the proof, we introduce a new time variable defined by $\tau_{k,n} := T - t_{n-1}$ and observe that, assuming we have already accomplished the bound

$$(4.89) \quad \sum_{n=1}^N \tau_{k,n} \|\xi_k\|_{L^2(I_n, H^{-1}(\Omega))}^2 \leq ck^2 \|w_T\|_{H^1(\Omega)}^2,$$

we obtain the required estimate by

$$\begin{aligned} \|\xi_k\|_{L^1(I, H^{-1}(\Omega))}^2 &\leq \sum_{n=1}^N k_n \tau_{k,n}^{-1} \sum_{n=1}^N \tau_{k,n} \|\xi_k\|_{L^2(I_n, H^{-1}(\Omega))}^2 \\ &\leq Ck^2 \left(\log \frac{T}{k} + 1 \right) \|w_T\|_{H^{-1}(\Omega)}^2. \end{aligned}$$

Therefore, we focus on the derivation of (4.89).

Testing the error equation for ξ_k , given by (4.82), with $\varphi_k := \tau_{k,n} \Delta^{-2} \xi_k$ we get

$$(4.90) \quad \begin{aligned} &(\tau_{k,n} \nabla \Delta^{-2} \xi_k, \nabla \xi_k)_{I_n} - (\tau_{k,n} \Delta^{-2} \xi_{k,n}, [\xi_k]_n) \\ &= \int_{I_n} (t_n - t) (\tau_{k,n} \Delta^{-1} \varepsilon_k, \partial_t w(t)) dt. \end{aligned}$$

For the first term in the left-hand side we apply twice integration by parts in space obtain

$$\begin{aligned} (\tau_{k,n} \nabla \Delta^{-2} \varepsilon_k, \nabla \varepsilon_k)_{I_n} &= -(\tau_{k,n} \Delta^{-1} \xi_k, \xi_k)_{I_n} = -(\tau_{k,n} \Delta^{-1} \xi_k, \Delta \Delta^{-1} \xi_k)_{I_n} \\ &= \tau_{k,n} \|\nabla \Delta^{-1} \xi_k\|_{I_n}^2, \end{aligned}$$

while, for the second term, equality (4.32) implies

$$\begin{aligned} -(\tau_{k,n} \Delta^{-2} \xi_{k,n}, [\xi_k]_n) &= -\tau_{k,n} (\Delta^{-1} \varepsilon_{k,n}, [\Delta^{-1} \varepsilon_k]_n) \\ &= \frac{1}{2} \tau_{k,n} (-\|\Delta^{-1} \xi_{k,n+1}\|^2 + \|\Delta^{-1} \xi_{k,n}\|^2 + \|[\Delta^{-1} \xi_k]_n\|^2) \\ &\geq \frac{1}{2} \tau_{k,n} (-\|\Delta^{-1} \xi_{k,n+1}\|^2 + \|\Delta^{-1} \xi_{k,n}\|^2) \end{aligned}$$

For the right-hand side of (4.90), we combine integration by parts in space and Young's inequality to write

$$\begin{aligned} &\int_{I_n} (t_n - t) (\tau_{k,n} \Delta^{-1} \xi_k, \partial_t w(t)) dt \\ &= - \int_{I_n} (t_n - t) (\tau_{k,n} \nabla \Delta^{-1} \xi_k, \nabla \Delta^{-1} \partial_t w(t)) dt \\ &\leq \frac{\tau_{k,n}}{2} \|\nabla \Delta^{-1} \xi_k\|_{I_n}^2 + \frac{\tau_{k,n}}{2} \int_{I_n} (t_n - t)^2 \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt. \end{aligned}$$

Inserting the previously derived relations in (4.90), we have

$$\begin{aligned} & \tau_{k,n} \|\nabla \Delta^{-1} \xi_k\|_{I_n}^2 + \tau_{k,n} \|\Delta^{-1} \xi_{k,n}\|^2 \\ & \leq \tau_{k,n} \|\Delta^{-1} \xi_{k,n+1}\|^2 + \tau_{k,n} \int_{I_n} (t_n - t)^2 \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt. \end{aligned}$$

We note that $\tau_{k,n} = \tau_{k,n+1} + k_n$ which, used in the term $\|\Delta^{-1} \xi_{k,n+1}\|^2$, lead to

$$(4.91) \quad \begin{aligned} & \tau_{k,n} \|\nabla \Delta^{-1} \xi_k\|_{I_n}^2 + \tau_{k,n} \|\Delta^{-1} \xi_{k,n}\|^2 - \tau_{k,n+1} \|\Delta^{-1} \xi_{k,n+1}\|^2 \\ & \leq k_n \|\Delta^{-1} \xi_{k,n+1}\|^2 + \tau_{k,n} \int_{I_n} (t_n - t)^2 \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt. \end{aligned}$$

Up to now, we have worked on a generic time interval I_n , while the relation to be shown, namely (4.89), is formulated on the whole time interval I . Therefore, we sum (4.91) over $n = 1, \dots, N$, use $\xi_{k,N+1} = 0$, and recall that the time-mesh satisfy $k_n \leq \tilde{k} k_{n+1}$, to obtain

$$(4.92) \quad \begin{aligned} & \sum_{n=1}^N \tau_{k,n} \|\nabla \Delta^{-1} \xi_k\|_{I_n}^2 + T \|\Delta^{-1} \xi_{k,1}\|^2 \\ & \leq \tilde{k} \|\Delta^{-1} \xi_k\|_I^2 + \sum_{n=1}^N \tau_{k,n} \int_{I_n} (t_n - t)^2 \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt. \end{aligned}$$

We note that the first term in the right-hand side of (4.92) is already bounded by the estimate (4.88), while the second term is similar to the time-weighted estimate (4.23a), which would lead to the sought bound, i.e., (4.89). To accomplish (4.23a), in a first step we estimate $(t_n - t)$ from above with k_n . Then, observing that

$$\tau_{k,n} \leq (1 + \tilde{k})(T - t), \text{ for } t \in I_n, n = 1, \dots, N - 1$$

and $\tau_{k,N} = k_N$, we split accordingly the second term in the right-hand side of (4.92) to conclude

$$(4.93) \quad \begin{aligned} & \sum_{n=1}^N \tau_{k,n} \int_{I_n} (t_n - t)^2 \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt \\ & \leq \sum_{n=1}^{N-1} k_n^2 \int_{I_n} \tau_{k,n} \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt + k_N^2 \int_{I_N} (T - t) \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt \\ & \leq (1 + \tilde{k}) k^2 \int_I (T - t) \|\nabla \Delta^{-1} \partial_t w(t)\|^2 dt \\ & \leq C k^2 \|w_T\|_{H^{-1}(\Omega)}, \end{aligned}$$

thanks to (4.23a).

Inserting the above bound in (4.92) together with the estimate (4.88) for $\|\Delta^{-1} \xi_k\|_I^2$, we obtain (4.89) as requested. \square

Last result has concluded the proof of Theorem 4.3.1 and we can now present the error estimate for e_k which has been derived in [56, Theorem 4.8] of the author of this thesis. The state space U used below is the one defined in (3.1) while the additional regularity of g and u_0 has been discussed in Remark 3.1.2.

Theorem 4.3.6. *Let $u \in U$ and $u_k \in U_k$ be solutions of (2.22) and (2.23), respectively, with $f(t, x) = q(t)g(x) \in L^\infty(I, H_0^1(\Omega))$ and $u_0 \in \dot{H}^3(\Omega)$. Then, for the error induced by the discretization in time it holds*

$$(4.94) \quad \|u - u_k\|_{L^\infty(I, H_0^1(\Omega))} \leq Ck \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \left(\|f\|_{L^\infty(I, H_0^1(\Omega))} + \|u_0\|_{\dot{H}^3(\Omega)} \right).$$

Proof. On each time interval I_n , $n = 1, \dots, N$, we split the error into the interpolation error and the error inside I_n , $n = 1, \dots, N$ writing

$$(4.95) \quad \|u - u_k\|_{L^\infty(I_n, H_0^1(\Omega))} \leq \|u(\cdot) - u(t_n)\|_{L^\infty(I_n, H_0^1(\Omega))} + \|u(t_n) - u_k(\cdot)\|_{L^\infty(I_n, H_0^1(\Omega))}.$$

One of the features of the discontinuous Galerkin method is the possibility to analyze quantities of interest on each time interval independently. We exploit this fact here considering the two terms in the right-hand side separately on each time interval I_n and then, summing over $n = 1, \dots, N$, we obtain the assertion.

Further, with no loss of generality, we focus on the last time interval I_N denoting by $\hat{t} \in I_N$ a generic fixed time. Indeed, to show the claim on a generic I_n , it is sufficient to use same arguments considering (4.1.1) on $I = (0, t_n)$ and (4.1.2) on $\hat{I} = (0, \hat{t})$ for $\hat{t} \in (t_{n-1}, t_n]$, observing that $0 \leq \log(t_n/k) \leq \log(T/k)$.

- (i) We start the analysis with the interpolation error $u(\hat{t}) - u(t_N)$ for which it will be employed the continuous estimate derived previously in Lemma 4.3.2. We consider the solutions w and \hat{w} of (4.3) and (4.4), respectively, with initial value w_T to be specified later. Firstly, we integrate by parts in time (4.3) and (4.4) getting

$$\begin{aligned} -(\varphi(T), w(T)) + (\varphi(0), w(0)) + (\partial_t \varphi, w)_I + (\nabla \varphi, \nabla w)_I &= 0, \\ -(\varphi(\hat{t}), \hat{w}(\hat{t})) + (\varphi(0), \hat{w}(0)) + (\partial_t \varphi, \hat{w})_{\hat{I}} + (\nabla \varphi, \nabla \hat{w})_{\hat{I}} &= 0, \end{aligned}$$

for any $\varphi \in U$.

Then, through the choice $\varphi = u$, we observe that the last two terms in the left-hand side, by means of the state equation (2.22), yield

$$(4.96) \quad -(u(T), w(T)) + (u(0), w(0)) + (f, w)_I = 0,$$

$$(4.97) \quad -(u(\hat{t}), \hat{w}(\hat{t})) + (u(0), \hat{w}(0)) + (f, \hat{w})_{\hat{I}} = 0.$$

By definition we have $w(T) = w(\hat{t}) = w_T$, therefore, subtracting (4.96) from (4.97) we have

$$(4.98) \quad (u(\hat{t}) - u(T), w_T) = (u(0), \hat{w}(0) - w(0)) + (f, \hat{w} - w)_{\hat{I}} - (f, w)_{I \setminus \hat{I}}.$$

We are now at the crucial point of the proof. The initial data w_T must be selected conveniently in order to obtain from the previous inequality the error in the correct norm, i.e., $H_0^1(\Omega)$. Further, such choice must also match with the $H^{-1}(\Omega)$ -norm, because, as already anticipated, we will make use of (4.63) where the bound depends on $\|w_T\|_{H^{-1}(\Omega)}$.

We observe that the choice $w_T = -\Delta(u(\hat{t}) - u(T))$ has the desired property, indeed, integration by parts in the left-hand side of (4.98) gives

$$\|\nabla(u(\hat{t}) - u(T))\|^2,$$

and

$$\|w_T\|_{H^{-1}(\Omega)} = \|\Delta^{-1}w_T\| = \|\nabla(u(\hat{t}) - u(T))\|.$$

With this in mind, we now employ the duality argument in (4.98) obtaining

$$\begin{aligned} \|\nabla(u(\hat{t}) - u(T))\|^2 &= (u(0), \hat{w}(0) - w(0)) + (f, \hat{w} - w)_{\hat{I}} - \int_{\hat{t}}^T (f(t), w(t)) dt \\ &\leq \left(\|\hat{w} - w\|_{L^1(\hat{I}, H^{-1}(\Omega))} + \|\hat{w}(0) - w(0)\|_{H^{-3}(\Omega)} \right. \\ &\quad \left. + k\|w\|_{L^\infty(I, H^{-1}(\Omega))} \right) \left(\|f\|_{L^\infty(I, H_0^1(\Omega))} + \|u_0\|_{\dot{H}^3(\Omega)} \right) \\ &\leq Ck \log\left(\frac{T}{k} + 1\right)^{\frac{1}{2}} \|w_T\|_{H^{-1}(\Omega)} (\|f\|_{L^\infty(I, H_0^1(\Omega))} + \|u_0\|_{\dot{H}^3(\Omega)}), \end{aligned}$$

where in the last inequality (3.7) and Lemma 4.3.2 has been used to bound $\|w\|_{L^\infty(I, H^{-1}(\Omega))}$ and the remaining terms, respectively.

Then, simplifying $\|w_T\|_{H^{-1}(\Omega)}$ with the left-hand side, we conclude

$$(4.99) \quad \|\nabla(u(\hat{t}) - u(T))\| \leq Ck \log\left(\frac{T}{k} + 1\right)^{\frac{1}{2}} (\|f\|_{L^\infty(I, H_0^1(\Omega))} + \|u_0\|_{\dot{H}^3(\Omega)}).$$

- (ii) For the error in the interior of the time interval, we proceed similarly to the previous case using now Problem 4.1.4(a). Therefore, we consider w and w_k solutions of (4.3) and (4.5), respectively, with initial value $w_T = -\Delta(u(t) - u(T))$. This choice gives

$$\begin{aligned} B(\varphi_k, w) &= B(\varphi_k, w_k) = (\varphi_{k,N}, -\Delta(u(T) - u_k(t))) \\ &= (\nabla\varphi_{k,N}, \nabla(u(T) - u_{k,N})) \end{aligned}$$

for any $\varphi_k \in U_k$.

In particular, through the choice $\varphi_k = u - u_k$, by means of Galerkin orthogonality, we have

$$\begin{aligned} \|\nabla(u(T) - u_{k,N})\|^2 &= B(u - u_k, w) = B(u - u_k, w - w_k) = B(u, w - w_k) \\ &= (f, w - w_k)_I + (u_0, w(0) - w_k(0)) \\ &\leq (\|w - w_k\|_{L^1(I, H^{-1}(\Omega))} + \|w(0) - w_k(0)\|_{H^{-3}(\Omega)}) \cdot \\ &\quad \cdot (\|f\|_{L^\infty(I, H_0^1(\Omega))} + \|u_0\|_{\dot{H}^3(\Omega)}). \end{aligned}$$

Then, Theorem 4.3.1 entails

$$(4.100) \quad \|\nabla(u(T) - u_{k,N})\| \leq Ck \log\left(\frac{T}{k} + 1\right)^{\frac{1}{2}} (\|f\|_{L^\infty(I, H_0^1(\Omega))} + \|u_0\|_{\dot{H}^3(\Omega)}).$$

In conclusion, combining (4.99) and (4.100) in (4.95), we have shown the desired estimate for the interval I_N . Proceeding similarly for all $I_n, n = 1, \dots, N-1$ and summing over n , the thesis follows. \square

4.3.2 Spatial error

The semi-discrete solution of Problem 4.1.4(a) and its discrete counterpart solution of Problem 4.1.4(b) are compared introducing the L^2 -projection in space P_h . Thus, the problem is again reduced into a projection error η_h and a discretization one ξ_h . The analysis is somehow simplified observing that the functions involved are piecewise constant in time, leading to L^2 -norm estimate in time instead of L^∞ -norm. The estimates required for the error in the dual variable ε_h are given in the following result. The proof is again given with a series of intermediates lemmas.

Theorem 4.3.7. *Let w_k and w_{kh} be solutions of (4.5) and (4.6), respectively. Then, the corresponding error satisfies*

$$(4.101) \quad \|w_k - w_{kh}\|_{L^2(I, H^{-1}(\Omega))} + \|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} \leq Ch \|w_T\|_{H^{-1}(\Omega)}$$

Proof. We split the error by means of the L^2 -projection P_h , observing that $P_h w_{kh} = w_{kh}$, writing

$$\begin{aligned} \|w_k - w_{kh}\|_{L^2(I, H^{-1}(\Omega))} + \|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} &\leq \|w_k - P_h w_k\|_{L^2(I, H^{-1}(\Omega))} \\ &\quad + \|P_h w_k - w_{kh}\|_{L^2(I, H^{-1}(\Omega))} + \|P_h w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)}. \end{aligned}$$

Then, the claim follows combining Lemmas 4.3.9, 4.3.10 and 4.3.12. \square

The first accessory result is relative to the error at the final time between the solutions v_k and v_{kh} of the auxiliary forward problems. This is needed to bound the error in the H^{-2} -norm.

Lemma 4.3.8. *For the error at the final time between the solutions v_k and v_{kh} of (4.7) and (4.8), respectively, there holds*

$$(4.102) \quad \|v_{k,N} - v_{kh,N}\| \leq Ch^2 \|\Delta v_0\|.$$

Proof. For the proof we refer to [61, Lemma 5.7(b)]. \square

Next lemma deals with the projection error η_h induced by P_h .

Lemma 4.3.9. *For the error between the solution w_k of (4.5) and its projection in space, there holds*

$$(4.103) \quad \|w_k - P_h w_k\|_{L^2(I, H^{-1}(\Omega))} \leq Ch \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. In a first step, we observe that, by definition of the L^2 -projection and standard estimates for it, for $\psi \in H_0^1(\Omega)$, we have

$$\begin{aligned} (w_k - P_h w_k, \psi) &= (w_k - P_h w_k, P_h \psi) + (w_k, \psi - P_h \psi) - (P_h w_k, \psi - P_h \psi) \\ &= (w_k, \psi - P_h \psi) \\ &\leq \|w_k\| \|\psi - P_h \psi\| \leq C \|w_k\| \|\psi - P_h \psi\|_{H_0^1(\Omega)} \\ &\leq Ch \|w_k\| \|\psi\|_{H_0^1(\Omega)}. \end{aligned}$$

We use the relation above in the definition of the H^{-1} -norm

$$\begin{aligned} \|w_k - P_h w_k\|_{H^{-1}(\Omega)} &= \sup_{\psi \in H_0^1(\Omega)} \frac{(w_k - P_h w_k, \psi)}{\|\psi\|_{H_0^1(\Omega)}} \\ &\leq Ch \|w_k\|, \end{aligned}$$

from which we deduce

$$(4.104) \quad \|w_k - P_h w_k\|_{L^2(I, H^{-1}(\Omega))} \leq Ch \|w_k\|_I \leq Ch \|w_T\|_{H^{-1}(\Omega)},$$

using (4.28). \square

We focus now on the $H^{-2}(\Omega)$ -norm estimate for the discretization error

$$\xi_h = P_h w_k - w_{kh}$$

at the initial time. This result corresponds to [56, Lemma 4.10] of the author of this thesis.

Lemma 4.3.10. *Let w_k, w_{kh} be solutions of (4.5) and (4.6), respectively. Then, there holds*

$$(4.105) \quad \|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} \leq Ch \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. We notice that

$$(4.106) \quad \|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} \simeq \sup_{\psi \in \dot{H}^2(\Omega)} \frac{(w_{k,1} - w_{kh,1}, \psi)}{\|\psi\|_{\dot{H}^2(\Omega)}}.$$

This observation suggests to provide an upper bound for the numerator in terms of $\|\psi\|_{H^2(\Omega)}$ and $\|w_T\|_{H^{-1}(\Omega)}$.

Before doing so, we have to derive an expression for $(w_{k,1} - w_{kh,1}, \psi)$. The idea is to pick the test functions in the auxiliary Problems 4.1.4 and 4.1.5 so that, at any level of discretization, the left-hand side of the backward problems coincides with the one of the forward.

Namely, for a fixed $\psi \in \dot{H}^2(\Omega)$, we consider $v_0 = \psi$ in (4.7) and (4.8). Then, in a first step we test (4.5) and (4.7) with $\varphi = v_k$ and $\varphi = w_k$, respectively. Afterwards, we test (4.6) and (4.8) with $\varphi = v_{kh}$ and $\varphi = w_{kh}$, respectively, obtaining

$$\begin{aligned} (\psi, w_{k,1}) &= B(v_k, w_k) = (v_{k,N}, w_T), \\ (\psi, w_{kh,1}) &= B(v_{kh}, w_{kh}) = (v_{kh,N}, w_T). \end{aligned}$$

Subtracting the two inequalities above we have

$$(4.107) \quad (\psi, w_{k,1} - w_{kh,1}) = B(v_k, w_k) - B(v_{kh}, w_{kh}) = (v_{k,N} - v_{kh,N}, w_T)$$

and we observe that the left-hand side is indeed what we need.

For the terms containing the bilinear form $B(\cdot, \cdot)$, we rearrange the terms as follow

$$\begin{aligned} B(v_k, w_k) - B(v_{kh}, w_{kh}) &= B(v_k - v_{kh}, w_k) + B(v_{kh}, w_k) \\ &\quad - B(v_{kh}, w_{kh} - w_k) - B(v_{kh}, w_k) \\ &= B(v_k - v_{kh}, w_k - w_{kh}) + B(v_k - v_{kh}, w_{kh}) \\ &\quad - B(v_{kh}, w_{kh} - w_k) \\ &= B(v_k - v_{kh}, w_k - w_{kh}), \end{aligned}$$

using Galerkin orthogonality in the last equality.
The above equality in (4.107) gives

$$(\psi, w_{k,1} - w_{kh,1}) = B(v_k - v_{kh}, w_k - w_{kh}) = (v_{k,N} - v_{kh,N}, w_T),$$

from which, after integration by parts,

$$(4.108) \quad \begin{aligned} (\psi, w_{k,1} - w_{kh,1}) &= -(\nabla(v_{k,N} - v_{kh,N}), \nabla \Delta^{-1} w_T) \\ &\leq \|\nabla(v_{k,N} - v_{kh,N})\| \|\nabla \Delta^{-1} w_T\|. \end{aligned}$$

The second term in the right-hand side already displays the $H^{-1}(\Omega)$ -norm of the initial data w_T , the same we are seeking in (4.105), hence we focus on the first term.

Denoting with \mathcal{I}_h the usual interpolation operator on V_h , we easily get

$$\|\nabla(v_{k,N} - v_{kh,N})\| \leq \|\nabla(v_{k,N} - \mathcal{I}_h v_{k,N})\| + \|\nabla(\mathcal{I}_h(v_{k,N}) - v_{kh,N})\|.$$

Then, recalling the following standard interpolation and inverse estimates for the case at hand

$$\begin{aligned} \|\nabla(v_{k,N} - \mathcal{I}_h v_{k,N})\| &\leq Ch \|\nabla^2 v_{k,N}\|, \\ \|\nabla(\mathcal{I}_h v_{k,N} - v_{kh,N})\| &\leq Ch^{-1} \|\mathcal{I}_h v_{k,N} - v_{kh,N}\|, \end{aligned}$$

we obtain

$$\begin{aligned} \|\nabla(v_{k,N} - v_{kh,N})\| &\leq Ch \|\nabla^2 v_{k,N}\| + ch^{-1} (\|\mathcal{I}_h v_{k,N} - v_{k,N}\| + \|v_{k,N} - v_{kh,N}\|) \\ &\leq C(h \|\Delta v_{k,N}\| + h^{-1} \|v_{k,N} - v_{kh,N}\|) \end{aligned}$$

where in the last estimates we used [39, Theorem 3.1.3.1].

Using Theorem 4.2.3 and Lemma 4.3.8 in the right-hand side, we assert

$$\|\nabla(v_{k,N} - v_{kh,N})\| \leq Ch(\|\Delta v_{k,N}\| + \|\Delta v_0\|) \leq Ch \|\Delta v_0\|,$$

which implies, back in (4.108),

$$\begin{aligned} (\psi, w_{k,1} - w_{kh,1}) &\leq Ch \|\Delta v_0\| \|\nabla \Delta^{-1} w_T\| \\ &\leq Ch \|v_0\|_{H^2(\Omega)} \|w_T\|_{H^{-1}(\Omega)}. \end{aligned}$$

Recalling that $\psi = v_0$, we can simplify the term $\|\psi\|_{H^2(\Omega)}$ in the right-hand side of (4.106) to finally conclude

$$\|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} \leq Ch \|w_T\|_{H^{-1}(\Omega)}.$$

□

In a last step, we write the equation for the discretization error ξ_h and bound it afterward. The following two results correspond to [56, Lemma 4.11].

Lemma 4.3.11. *The discretization error is formulated through the expression*

$$(4.109) \quad (\nabla \varphi, \nabla \xi_h)_{I_n} - (\varphi_n, [\xi_h]_n) = (\nabla \varphi, \nabla(P_h w_k - w_k))_{I_n},$$

for any $\varphi \in \mathcal{P}_0(I_n, V_h)$, $n = 1, \dots, N-1$.

Proof. We recall that $\varepsilon_h = w_k - w_{kh}$. Subtracting (4.6) to (4.5) and integrating by parts in time, we have

$$(4.110) \quad (\nabla \varphi, \nabla \varepsilon_h)_{I_n} - (\varphi_n, [\varepsilon_h]_n) = 0, \quad \forall \varphi \in \mathcal{P}_0(I_n, V_h), \quad n = 1, \dots, N-1.$$

Then, observing that $\varepsilon_h = w_k - P_h w_k + P_h \varepsilon_h$, and thanks to the definition of P_h , we have, for any $\varphi \in \mathcal{P}_0(I_n, V_h)$, $n = 1, \dots, N-1$

$$\begin{aligned} 0 &= (\nabla \varphi, \nabla (w_k - P_h w_k + P_h \varepsilon_h))_{I_n} - (\varphi_n, [w_k - P_h w_k + P_h \varepsilon_h]_n) \\ &= (\nabla \varphi, \nabla P_h \varepsilon_h)_{I_n} - (\varphi_n, [P_h \varepsilon_h]_n) - (\nabla \varphi, \nabla P_h w_k - w_k)_{I_n}, \end{aligned}$$

which gives the claim observing that $P_h \varepsilon_h$ coincides with ξ_h . \square

Lemma 4.3.12. *Let w_k and w_{kh} be solutions of (4.5) and (4.6), respectively. Then, for the discretization error there hold*

$$(4.111) \quad \|P_h w_k - w_{kh}\|_{L^2(I, H^{-1}(\Omega))} \leq Ch \|w_T\|_{H^{-1}(\Omega)}.$$

Proof. Recalling $\xi_h = P_h \varepsilon_h$, we select the test function in (4.109) so that we obtain $P_h \varepsilon_h$ in the correct norm. To this end, setting $\varphi = \Delta_h^{-2} P_h \varepsilon_h$ in (4.109), we get

$$(4.112) \quad \begin{aligned} &(\nabla \Delta_h^{-2} P_h \varepsilon_h, \nabla P_h \varepsilon_h)_{I_n} - (\Delta_h^{-2} P_h \varepsilon_h, [P_h \varepsilon_h]_n) \\ &= (\nabla \Delta_h^{-2} P_h \varepsilon_h, \nabla (P_h w_k - w_k))_{I_n}. \end{aligned}$$

We now analyze the three terms separately starting with the first one in the left-hand side. Using twice integrations by parts in space, we have

$$(4.113) \quad \begin{aligned} (\nabla \Delta_h^{-2} P_h \varepsilon_h, \nabla P_h \varepsilon_h)_{I_n} &= -(\Delta_h^{-1} P_h \varepsilon_h, P_h \varepsilon_h)_{I_n} \\ &= -(\Delta_h^{-1} P_h \varepsilon_h, \Delta_h \Delta_h^{-1} P_h \varepsilon_h)_{I_n} \\ &= \|\nabla \Delta_h^{-1} P_h \varepsilon_h\|_{I_n}^2, \end{aligned}$$

which is indeed the norm we are seeking for.

In the second term we employ relation (4.32) to write

$$(4.114) \quad \begin{aligned} -(\Delta_h^{-2} P_h \varepsilon_h, [P_h \varepsilon_h]_n) &= -(\Delta_h^{-1} P_h \varepsilon_h, [\Delta_h^{-1} P_h \varepsilon_h]_n) \\ &= \frac{1}{2} (\|\Delta_h^{-1} P_h \varepsilon_h\|^2 + \|[\Delta_h^{-1} P_h \varepsilon_h]_n\|^2) \\ &\quad - \|\Delta_h^{-1} P_h \varepsilon_h\|_{n+1}^2 \\ &\geq \frac{1}{2} (\|\Delta_h^{-1} P_h \varepsilon_h\|^2 - \|\Delta_h^{-1} P_h \varepsilon_h\|_{n+1}^2). \end{aligned}$$

For the right-hand side of (4.112), initially we proceed as in (4.113) integrating by parts twice, then, we use Young's inequality to obtain

$$(4.115) \quad \begin{aligned} (\nabla \Delta_h^{-2} P_h \varepsilon_h, \nabla (P_h w_k - w_k))_{I_n} &= -(\Delta_h^{-1} P_h \varepsilon_h, P_h w_k - w_k)_{I_n} \\ &= (\nabla \Delta_h^{-1} P_h \varepsilon_h, \nabla \Delta_h^{-1} (P_h w_k - w_k))_{I_n} \\ &\leq \frac{1}{2} (\|\nabla \Delta_h^{-1} P_h \varepsilon_h\|_{I_n}^2 \\ &\quad + \|\nabla \Delta_h^{-1} (P_h w_k - w_k)\|_{I_n}^2) \end{aligned}$$

Combining (4.113) with (4.114) and (4.115) in (4.112), we have

$$\|\nabla \Delta_h^{-1} P_h \varepsilon_h\|_{I_n}^2 + \|\Delta_h^{-1} P_h \varepsilon_{h,n}\|^2 - \|\Delta_h^{-1} P_h \varepsilon_{h,n+1}\|^2 \leq \|\nabla \Delta_h^{-1} (P_h w_k - w_k)\|_{I_n}^2$$

Adding these inequalities for $n = 1, \dots, N-1$, we obtain

$$\begin{aligned} \sum_{n=1}^{N-1} \|\nabla \Delta_h^{-1} P_h \varepsilon_h\|_{I_n}^2 + \|\Delta_h^{-1} P_h \varepsilon_{h,1}\|^2 - \|\Delta_h^{-1} P_h \varepsilon_{h,N}\|^2 \\ \leq \sum_{n=1}^{N-1} \|\nabla \Delta_h^{-1} (P_h w_k - w_k)\|_{I_n}^2. \end{aligned}$$

We observe that we can extend both sums up to the last time interval. Indeed, for the one in the left-hand side, we exploit $P_h \varepsilon_{h,N} = 0$, while in the right-hand side we are adding a non-negative term. Then, observing that the second term in the left-hand side is positive and the third is zero, we obtain

$$\|\nabla \Delta_h^{-1} P_h \varepsilon_h\|_I^2 \leq \|\nabla \Delta_h^{-1} (P_h w_k - w_k)\|_I^2,$$

and, recalling the equivalence between the discrete negative norm $\|\nabla \Delta_h^{-1} \cdot\|$ and the continuous one $\|\nabla \Delta^{-1} \cdot\|$, we get from (4.104) and (4.28)

$$\|\nabla \Delta^{-1} P_h \varepsilon_h\|_I^2 \leq \|\nabla \Delta^{-1} (P_h w_k - w_k)\|_I^2 \leq ch^2 \|w_k\|_I^2 \leq ch^2 \|w_T\|_{H^{-1}(\Omega)}$$

which concludes the proof. \square

Now that Theorem 4.3.7 has been derived, we conclude the section with the discretization error e_h , see [56, Theorem 4.12] of the thesis author.

Theorem 4.3.13. *Let $u_k \in U_k$ and $u_{kh} \in U_{kh}$ be solutions of (2.23) and (2.23), respectively, with $f(t, x) = q(t)g(x) \in L^\infty(I, H_0^1(\Omega))$ and $u_0 \in \dot{H}^2(\Omega)$. Then, for the error induced by the discretization in space it holds*

$$(4.116) \quad \|u_{kh} - u_k\|_{L^\infty(I, H_0^1(\Omega))} \leq Ch(\|f\|_{L^2(I, H_0^1(\Omega))} + \|u_0\|_{H^2(\Omega)}).$$

Proof. We observe that both u_k, u_{kh} are constant on I_n for any $n = 1, \dots, N$. Hence we can equivalently show the claim on I_n and with no loss of generality we consider the last time interval I_N only. We consider $w_k \in U_k$, $w_{kh} \in U_{kh}$ solutions of (4.5) and (4.6), respectively, with

$$w_T = -\Delta_h(u_{k,N} - u_{kh,N}).$$

Then, we employ the duality argument and exploit Galerkin orthogonality to get

$$\begin{aligned} \|\nabla(u_{k,N} - u_{kh,N})\|^2 &= B(u_k - u_{kh}, w_k) = B(u_k, w_k - w_{kh}) \\ &= (f, w_k - w_{kh})_I + (u_0, w_{k,1} - w_{kh,1}) \\ &\leq (\|w_k - w_{kh}\|_{L^2(I, H^{-1}(\Omega))} + \|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)}) \\ &\quad \cdot (\|f\|_{L^2(I, H_0^1(\Omega))} + \|u_0\|_{H^2(\Omega)}). \end{aligned}$$

We now use Theorem 4.3.7 to conclude

$$\|\nabla(u_{k,N} - u_{kh,N})\|^2 \leq ch \|w_T\|_{H^{-1}(\Omega)} (\|f\|_{L^2(I, H_0^1(\Omega))} + \|u_0\|_{H^2(\Omega)}).$$

Then, the assertion follows observing that by our choice of w_T it holds

$$\|w_T\|_{H^{-1}(\Omega)} = \|\nabla(u_{k,N} - u_{kh,N})\|.$$

\square

4.3. *Error analysis for linear equations*

Combining Theorems 4.3.6 and 4.3.13, we obtain the total discretization error formulated in (4.1).

4.4 Error Analysis for semi-linear equations

We now move our attention to the semi-linear state equation (2.28) deriving $L^\infty(I, L^2(\Omega))$ -norm estimates for the error arising from the discretization. In order to deal with the presence of the semi-linear term $d(\cdot, \cdot, u)$, we will need error estimates in the $L^2(I, L^2(\Omega))$ -norm for the solutions of the auxiliary linearized problems defined in Section 4.1.

4.4.1 Temporal error

We start the analysis deriving an $L^2(I, L^2(\Omega))$ -norm estimate for the solutions of the continuous auxiliary Problems 4.1.6(a)-(b), which will be used later in Lemma 4.4.5.

Theorem 4.4.1. *Let w be the solution of (4.11) and \hat{w} be the solution of (4.12). Then for the corresponding error there holds*

$$(4.117) \quad \|w - \hat{w}\|_{L^2(\hat{I}, L^2(\Omega))} \leq Ck \|w_T\|.$$

Proof. We denote $\hat{e} = w - \hat{w}$ and introduce an auxiliary backward problem with \hat{e} as data in the right-hand side. Find \bar{w} solution of

$$(4.118) \quad B(\varphi_k, \bar{w}) + (\varphi_k, \tilde{d}\bar{w})_{\hat{I}} = (\varphi_k, \hat{e})_{\hat{I}}$$

for any $\varphi_k \in U_k$.

Since the solutions w and \hat{w} possess the regularity $W(0, T) \hookrightarrow C(I, L^2(\Omega))$, the problem above is well-defined and, further, w and \hat{w} satisfy respectively the semi-discrete equation

$$\begin{aligned} B(\varphi_k, w) &= (\varphi_k, w_T) - (\varphi_k, \tilde{d}w)_I \\ B(\varphi_k, \hat{w}) &= (\varphi_k, w_{T-1}) - (\varphi_k, \tilde{d}\hat{w})_{\hat{I}}, \end{aligned}$$

for any $\varphi_k \in U_k$. Subtracting the two equations above, we obtain

$$(4.119) \quad B(\varphi_k, \hat{e}) = -(\varphi_k, \tilde{d}\hat{e})_{\hat{I}},$$

which will be useful in the sequel.

In a second step, we split the error as

$$\begin{aligned} \hat{e} &= (w - \pi_k w) + (\pi_k w - \pi_k \hat{w}) + (\pi_k \hat{w} - \hat{w}) \\ &= \eta_k + \xi_k + \hat{\eta}_k, \end{aligned}$$

and rewrite (4.119) as

$$(4.120) \quad B(\varphi_k, \xi_k) = -(\varphi_k, \tilde{d}\hat{e})_{\hat{I}} - B(\varphi_k, \eta_k + \hat{\eta}_k).$$

We now test (4.118) with $\varphi_k = \xi_k$ and, thanks to (4.120), we have

$$\begin{aligned} (\xi_k, \hat{e})_{\hat{I}} &= B(\xi_k, \bar{w}) + (\xi_k, \tilde{d}\bar{w})_{\hat{I}} \\ &= -(\bar{w}, \tilde{d}\hat{e})_{\hat{I}} - B(\bar{w}, \eta_k + \hat{\eta}_k) + (\xi_k, \tilde{d}\bar{w})_{\hat{I}} \\ &= -B(\bar{w}, \eta_k + \hat{\eta}_k) - (\eta_k + \hat{\eta}_k, \tilde{d}\bar{w})_{\hat{I}}. \end{aligned}$$

Using again the splitting, this implies

$$\begin{aligned}\|\hat{e}\|_{\hat{I}}^2 &= -(\eta_k + \hat{\eta}_k, \hat{e})_{\hat{I}} - B(\bar{w}, \eta_k + \hat{\eta}_k) - (\eta_k + \hat{\eta}_k, \tilde{d}\bar{w})_{\hat{I}} \\ &\leq \left(\|\eta_k\|_{\hat{I}} + \|\hat{\eta}_k\|_{\hat{I}}\right)\|\hat{e}\|_{\hat{I}} + \left(\|\eta_k\|_{\hat{I}} + \|\hat{\eta}_k\|_{\hat{I}}\right)\|\Delta\bar{w}\|_{\hat{I}} \\ &\quad + \|\tilde{d}\|_{\infty, \infty} \left(\|\eta_k\|_{\hat{I}} + \|\hat{\eta}_k\|_{\hat{I}}\right)\|\bar{w}\|_{\hat{I}}\end{aligned}$$

Thanks to the boundedness in $L^\infty(I \times \Omega)$ of \tilde{d} , the solution \bar{w} of (4.118) exhibits the regularity

$$(4.121) \quad \|\bar{w}\|_{\hat{I}} + \|\Delta\bar{w}\|_{\hat{I}} \leq C\|\hat{e}\|_{\hat{I}},$$

compare with [68, Corollary 3.2]. Then, exploiting the inequality above and the approximation property of π_k , we obtain

$$\begin{aligned}\|\hat{e}\|_{\hat{I}}^2 &\leq C\left(\|\eta_k\|_{\hat{I}} + \|\hat{\eta}_k\|_{\hat{I}}\right)\|\hat{e}\|_{\hat{I}} \\ &\leq Ck\left(\|\partial_t w\|_{\hat{I}} + \|\partial_t \hat{w}\|_{\hat{I}}\right)\|\hat{e}\|_{\hat{I}} \\ &\leq Ck\|w_T\|\|\hat{e}\|_{\hat{I}},\end{aligned}$$

where in the last step we have used (4.34a). Then, the claim follows dividing the inequality above by $\|\hat{e}\|_{\hat{I}}$. \square

We require a similar estimate as the one before for the error between the solution of Problem 4.1.6(a) and the solution of Problem 4.1.7(a). It will be used to treat the discretization error in Lemma 4.4.8.

Theorem 4.4.2. *Let w be the solution of (4.11) and w_k be the solution of (4.13) defined through \tilde{d} . Then, for the corresponding error there holds*

$$(4.122) \quad \|w - w_k\|_{L^2(I, L^2(\Omega))} \leq Ck\|w_T\|.$$

Proof. We define an auxiliary problem having as data the error ε_k

$$(4.123) \quad B(\bar{v}_k, \varphi_k) + (\tilde{d}\bar{v}_k, \varphi_k)_I = (\varepsilon_k, \varphi_k)_I, \quad \forall \varphi_k \in U_k,$$

whose regularity

$$(4.124) \quad \|\bar{v}_k\|_I + \|\Delta\bar{v}_k\|_I \leq C\|\varepsilon_k\|_I$$

follows as (4.121) thanks to the boundedness of \tilde{d} .

Before selecting conveniently the equation above, we observe that the definition of π_k entails

$$(4.125) \quad B(\eta_k, \varphi_k) = (\nabla\eta_k, \nabla\varphi_k)_I, \quad \forall \varphi_k \in U_k,$$

see [62, Lemma 5.2].

Then, testing (4.123) with $\varphi_k = \xi_k$ and thanks to (4.125), we have

$$\begin{aligned}(\varepsilon_k, \xi_k)_I &= B(\bar{v}_k, \xi_k) + (\tilde{d}\bar{v}_k, \xi_k)_I \\ &= B(\bar{v}_k, \varepsilon_k - \eta_k) + (\tilde{d}\bar{v}_k, \varepsilon_k - \eta_k)_I \\ &= B(\bar{v}_k, \varepsilon_k) - (\nabla\bar{v}_k, \nabla\eta_k)_I + (\tilde{d}\bar{v}_k, \varepsilon_k - \eta_k)_I \\ &= -(\bar{v}_k, \tilde{d}\varepsilon_k)_I - (\nabla\bar{v}_k, \nabla\eta_k)_I + (\tilde{d}\bar{v}_k, \varepsilon_k - \eta_k)_I,\end{aligned}$$

using (4.19) in the last step.

Using again $\xi_k = \varepsilon_k - \eta_k$, integration by parts and (4.124) we conclude

$$\begin{aligned}
 \|\varepsilon_k\|_I^2 &= (\varepsilon_k, \eta_k)_I + (\Delta \bar{v}_k, \eta_k)_I - (\tilde{d} \bar{v}_k, \eta_k)_I \\
 &\leq \left(\|\varepsilon_k\| + \|\Delta \bar{v}_k\|_I + \|\tilde{d}\|_{\infty, \infty} \|\bar{v}_k\|_I \right) \|\eta_k\|_I \\
 &\leq C \|\varepsilon_k\|_I \|\eta_k\|_I \\
 &\leq Ck \|\varepsilon_k\|_I \|\partial_t w\|_I \\
 &\leq Ck \|\varepsilon_k\|_I \|w_T\|,
 \end{aligned}$$

and the claim follows dividing by $\|\varepsilon_k\|_I$. \square

Remark 4.4.3. In the context of the L^∞ -norm estimate in time, the projector π_k has been defined on $I \setminus I_N$ to exploit the error arising by the truncation of the equation. However, π_k can be equivalently defined on the whole time interval I , see, e.g., [84, Equation (12.9)].

After this preparation, the exposition follows now as in the previous section. In a first step, we state the main result regarding the error arising from the time discretization of the auxiliary problem defined by (4.11).

Theorem 4.4.4. Let w be the solution of (4.11) and w_k be the solution of (4.13) defined through \tilde{d} . Then, the corresponding error satisfies

$$(4.126) \quad \|w - w_k\|_{L^1(I, L^2(\Omega))} + \|w(0) - w_{k,1}\|_{H^{-2}(\Omega)} \leq Ck \left(\log \frac{T}{k} + 1 \right) \|w_T\|.$$

Proof. We use the approach outlined in Theorem 4.3.1 introducing the projection operator π_k . The error is divided in three parts

$$\begin{aligned}
 \|w - w_k\|_{L^1(I, L^2(\Omega))} + \|w(0) - w_{k,1}\|_{H^{-2}(\Omega)} &\leq \|w - \pi_k w\|_{L^1(I, L^2(\Omega))} \\
 &\quad + \|\pi_k w - w_k\|_{L^1(I, L^2(\Omega))} + \|\pi_k w(0) - w_{k,1}\|_{H^{-2}(\Omega)},
 \end{aligned}$$

and the claim follows by Lemmas 4.4.6 and 4.4.8. \square

Lemma 4.4.5. For the error between w and \hat{w} solutions of (4.11) and (4.12), respectively, there holds

$$(4.127) \quad \|w - \hat{w}\|_{L^1(\hat{I}, L^2(\Omega))} + \|w(0) - \hat{w}(0)\|_{H^{-2}(\Omega)} \leq Ck \left(\log \frac{T}{k} + 1 \right) \|w_T\|.$$

Proof. Denoting the error with $\hat{\varepsilon} := w - \hat{w}$ and proceeding as in Lemma 4.3.2, the equation for $\hat{\varepsilon}$ reads

$$(4.128) \quad -(\varphi, \partial_t \hat{\varepsilon})_{\hat{I}} + (\nabla \varphi, \nabla \hat{\varepsilon})_{\hat{I}} + (\varphi, \tilde{d} \hat{\varepsilon})_{\hat{I}} = 0,$$

for any $\varphi \in W(0, T)$.

We now analyze separately the two terms in (4.127) starting with the $H^{-2}(\Omega)$ -norm

(i) We select $\varphi = \Delta^{-2} \hat{\varepsilon}$ in (4.128) obtaining

$$(4.129) \quad -(\Delta^{-2} \hat{\varepsilon}, \partial_t \hat{\varepsilon})_{\hat{I}} + (\nabla \Delta^{-2} \hat{\varepsilon}, \nabla \hat{\varepsilon})_{\hat{I}} + (\Delta^{-2} \hat{\varepsilon}, \tilde{d} \hat{\varepsilon})_{\hat{I}} = 0.$$

Using in the first term same remark as in (4.66), and observing that $\hat{\varepsilon}(\hat{t}) = w_T - w(\hat{t})$, we get

$$\begin{aligned} -(\Delta^{-2}\hat{\varepsilon}, \partial_t\hat{\varepsilon})_{\hat{I}} &= -(\Delta^{-2}\hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})) + (\Delta^{-2}\hat{\varepsilon}(0), \hat{\varepsilon}(0)) + (\partial_t(\Delta^{-2}\hat{\varepsilon}), \hat{\varepsilon})_{\hat{I}} \\ &= -\|\Delta^{-1}\hat{\varepsilon}(\hat{t})\|^2 + \|\Delta^{-1}\hat{\varepsilon}(0)\|^2 + \|\nabla\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2, \end{aligned}$$

where in the last term we have first used $\partial_t\hat{\varepsilon} = -\Delta\hat{\varepsilon}$ and then integration by part in space.

Back in (4.129), this yields

$$(4.130) \quad \|\Delta^{-1}\hat{\varepsilon}(0)\|^2 + 2\|\nabla\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2 = \|\Delta^{-1}\hat{\varepsilon}(\hat{t})\|^2 - (\Delta^{-2}\hat{\varepsilon}, \tilde{d}\hat{\varepsilon})_{\hat{I}}$$

We focus on the term containing \tilde{d} and, by means of Young's inequality and the compactness of the operator Δ^{-2} , we have

$$\begin{aligned} -(\Delta^{-2}\hat{\varepsilon}, \tilde{d}\hat{\varepsilon})_{\hat{I}} &\leq \|\tilde{d}\|_{\infty, \infty} \frac{1}{2} \left(\|\Delta^{-2}\hat{\varepsilon}\|_{\hat{I}}^2 + \|\hat{\varepsilon}\|_{\hat{I}}^2 \right) \\ (4.131) \quad &\leq \|\tilde{d}\|_{\infty, \infty} \left(c\|\hat{\varepsilon}\|_{\hat{I}}^2 \right) \\ &\leq Ck^2\|w_T\|^2, \end{aligned}$$

using Theorem 4.4.1 in the last step.

Concerning the first term in the right-hand side of (4.130), we have

$$\begin{aligned} \|\Delta^{-1}\hat{\varepsilon}(\hat{t})\|^2 &= \int_{\Omega} \left(\int_{\hat{t}}^T \Delta^{-1}\partial_t w(t) dt \right)^2 dx \\ (4.132) \quad &= \int_{\Omega} \left(\int_{\hat{t}}^T 1(\Delta^{-1}\partial_t w(t)) dt \right)^2 dx \\ &\leq k_N \int_{\hat{t}}^T \|w(t)\|^2 dt \\ &\leq Ck^2\|w_T\|^2 \end{aligned}$$

using similar argument as in (4.72). Then, the first part is concluded with

$$(4.133) \quad \|\Delta^{-1}\hat{\varepsilon}(0)\|^2 + 2\|\nabla\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2 \leq Ck^2\|w_T\|^2.$$

- (ii) As in (4.75), the claim follows once we show that for $\tau = \max(\hat{t} - t, k)$ it holds

$$(4.134) \quad \|\sqrt{\tau}\hat{\varepsilon}\|_{\hat{I}}^2 \leq Ck^2\|w_T\|^2.$$

To this end, we test (4.128) with $\varphi = -\tau\Delta^{-1}\hat{\varepsilon}$ obtaining

$$(4.135) \quad (\tau\Delta^{-1}\hat{\varepsilon}, \partial_t\hat{\varepsilon})_{\hat{I}} - (\tau\nabla\Delta^{-1}\hat{\varepsilon}, \nabla\hat{\varepsilon})_{\hat{I}} - (\tau\Delta^{-1}\hat{\varepsilon}, \tilde{d}\hat{\varepsilon})_{\hat{I}} = 0.$$

For the first term, we use (4.66), integration by parts and the relation $\partial_t\hat{\varepsilon} = -\Delta\hat{\varepsilon}$, to get

$$\begin{aligned} (\tau\Delta^{-1}\hat{\varepsilon}, \partial_t\hat{\varepsilon})_{\hat{I}} &= -(\partial_t(\tau\Delta^{-1}\hat{\varepsilon}), \hat{\varepsilon})_{\hat{I}} + \int_{\hat{I}} \partial_t(\tau\Delta^{-1}\hat{\varepsilon}, \hat{\varepsilon}) dt \\ (4.136) \quad &= -(\tau'\Delta^{-1}\hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} - (\tau\partial_t\Delta^{-1}\hat{\varepsilon}, \hat{\varepsilon})_{\hat{I}} \\ &\quad - k(\Delta^{-1}\hat{\varepsilon}(\hat{t}), \hat{\varepsilon}(\hat{t})) + \hat{t}\|\nabla\Delta^{-1}\hat{\varepsilon}(0)\|^2 \\ &= \tau'\|\nabla\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2 + \|\sqrt{\tau}\hat{\varepsilon}\|_{\hat{I}}^2 \\ &\quad - k(\Delta^{-1}\hat{\varepsilon}(\hat{t}), \hat{\varepsilon}(\hat{t})) + \hat{t}\|\nabla\Delta^{-1}\hat{\varepsilon}(0)\|^2. \end{aligned}$$

Then, we apply integration by parts to the second term in (4.135), insert the relation above, exploit the compactness of the inverse Laplacian and note that $-\tau' \leq 1$, obtaining

$$\begin{aligned}
 (4.137) \quad & 2\|\sqrt{\tau}\hat{\varepsilon}\|_{\hat{I}}^2 + \hat{t}\|\nabla\Delta^{-1}\hat{\varepsilon}(0)\|^2 = -\tau'\|\nabla\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2 \\
 & \quad k(\Delta^{-1}\hat{\varepsilon}(\hat{t}), w_T - w(\hat{t})) + (\tau\Delta^{-1}\hat{\varepsilon}, \tilde{d}\hat{\varepsilon})_{\hat{I}} \\
 & \leq \|\nabla\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2 + k\|\Delta^{-1}\hat{\varepsilon}(\hat{t})\|\|w_T\| + T\|\Delta^{-1}\hat{\varepsilon}\|_I\|\tilde{d}\hat{\varepsilon}\|_I \\
 & \leq \|\nabla\Delta^{-1}\hat{\varepsilon}\|_{\hat{I}}^2 + Ck\|\hat{\varepsilon}(\hat{t})\|\|w_T\| + C\|\hat{\varepsilon}(\hat{t})\|_{\hat{I}}^2.
 \end{aligned}$$

The first term in the right-hand side of (4.137) is bounded by (4.133), while for the remaining two terms we use Theorem 4.4.1 to obtain

$$(4.138) \quad 2\|\sqrt{\tau}\hat{\varepsilon}\|_{\hat{I}}^2 + \hat{t}\|\nabla\Delta^{-1}\hat{\varepsilon}(0)\|^2 \leq Ck^2\|w_T\|^2,$$

as requested. \square

In a next step, we bound the projection error exploiting the time-weighted estimates given in Proposition 4.2.4.

Lemma 4.4.6. *For the error between the solution w of (4.11) and its projection in time, there holds*

$$(4.139) \quad \|w - \pi_k w\|_{L^1(I, L^2(\Omega))} \leq Ck \left(\log \frac{T}{k} \right)^{\frac{1}{2}} \|w_T\|$$

Proof. The statement follows from the approximation property of π_k , see [84, Equation (12.10)], and the stability estimates (4.34a) and (4.34c). Indeed

$$\begin{aligned}
 \|w - \pi_k w\|_{L^1(I, L^2(\Omega))} &= \int_{I \setminus I_N} \|w - \pi_k w\| dt + \int_{I_N} \|w - \pi_k w\| dt \\
 &\leq Ck \left(\int_{I \setminus I_N} \|\partial_t w(t)\| dt + \max_{t \in I_N} \|w_t\| \right) \\
 &\leq Ck \left(\log \frac{T}{k} \right)^{\frac{1}{2}} \|w_T\|.
 \end{aligned}$$

\square

We are left with the discretization error $\xi_k = \pi_k w - w_k$ which will be bounded with the following two lemmas.

Lemma 4.4.7. *The discretization error ξ_k is formulated through the equation*

$$\begin{aligned}
 (4.140) \quad & (\nabla\varphi_k, \nabla\xi_k)_{I_n} - (\varphi_{k,n}, [\xi_k]_n) + (\varphi_k, \tilde{d}\xi_k)_{I_n} = \\
 & \int_{I_n} (t_n - t)(\Delta\varphi_k, \partial_t w(t)) dt + (\varphi_k, \tilde{d}(\pi_k w - w))_{I_n},
 \end{aligned}$$

for any $i = 1, \dots, N$ and $\varphi_k \in \mathcal{P}_0(I_n, H^2(\Omega) \cap H_0^1(\Omega))$.

Proof. The claim follows as in Lemma 4.3.4 with trivial changes due to the presence of the linearization term, which comes into play through (4.19). \square

Lemma 4.4.8. *Let w be the solution of (4.11) and w_k be the solution of 4.13 defined through \tilde{d} . Then for the discretization error there holds*

$$(4.141) \quad \|\pi_k w(0) - w_{k,1}\|_{H^{-2}(\Omega)} + \|\pi_k w - w_k\|_{L^1(I, L^2(\Omega))} \leq Ck \left(\log \frac{T}{k} + 1 \right) \|w_T\|.$$

Proof. We use similar arguments as in the proof of Lemma 4.3.5. Therefore, when no confusion arises, we omit the details and focus mostly on the difficulties introduced by the presence of the term containing \tilde{d} . The interested reader can refer also to [61, Lemma 5.2] where the claim is shown for the linear case.

To derive the estimate in the $H^{-2}(\Omega)$ -norm we test (4.140) with $\varphi = \Delta^{-2}(\xi_k)$ and, proceeding similarly as in (4.84)-(4.87), we have

$$(4.142) \quad \frac{1}{2} \left(\|\Delta^{-1} \xi_{k,n}\|^2 + \|\nabla \Delta^{-1} \xi\|_{I_n}^2 \right) \leq \frac{1}{2} \left(\|\Delta^{-1} \xi_{k,n+1}\|^2 + k_n^2 \|\nabla w\|_{I_n}^2 \right) + (\Delta^{-2} \xi_k, \tilde{d}(\pi_k w - w))_{I_n} - (\Delta^{-2} \xi_k, \tilde{d} \xi_k)_{I_n}.$$

We move our attention to the last two terms in the right-hand side of (4.142). Recalling the splitting (4.21), using Young's and Minkowski's inequality, and the compactness of Δ^{-2} , we get

$$\begin{aligned} (\Delta^{-2} \xi_k, \tilde{d}(\pi_k w - w))_{I_n} - (\Delta^{-2} \xi_k, \tilde{d} \xi_k)_{I_n} &= -(\Delta^{-2} \xi_k, \tilde{d} \varepsilon_k)_{I_n} \\ &\leq \frac{1}{2} (\|\Delta^{-2} \xi_k\|_{I_n}^2 + c \|\varepsilon_k\|_{I_n}^2) \\ &\leq \frac{C}{2} (\|\xi_k\|_{I_n}^2 + \|\varepsilon_k\|_{I_n}^2) \\ &\leq \frac{C}{2} (\|\eta_k\|_{I_n}^2 + 2 \|\varepsilon_k\|_{I_n}^2) \\ &\leq \frac{C}{2} (k_n^2 \|\partial_t w\|_{I_n}^2 + \|\varepsilon_k\|_{I_n}^2). \end{aligned}$$

Going back to (4.142) and summing for $n = 1, \dots, N$, we obtain

$$(4.143) \quad \|\nabla \Delta^{-1} \xi_k\|_I^2 + \|\Delta^{-1} \xi_{k,1}\|^2 \leq k^2 \|\nabla w\|_I^2 + C(\|\varepsilon_k\|_I^2 + k^2 \|\partial_t w\|_I^2) \leq Ck^2 \|w_T\|^2,$$

using the approximation property of π_k stated in (4.20), Proposition 4.2.4 and Theorem 4.4.2. This concludes the first part

For the $L^1(I, L^2(\Omega))$ -norm estimate, we observe that it is enough to show

$$(4.144) \quad \sum_{n=1}^N \tau_{k,n} \|\xi_k\|_{I_n}^2 \leq Ck^2 \|w_T\|^2, \quad \tau_{k,n} = T - t_{n-1},$$

to conclude

$$(4.145) \quad \begin{aligned} \|\xi_k\|_{L^1(I, L^2(\Omega))}^2 &\leq \sum_{n=1}^N k_n \tau_{k,n}^{-1} \sum_{n=1}^N \tau_{k,n} \|\xi_k\|_{I_n}^2 \\ &\leq Ck^2 \left(\log \frac{T}{k} + 1 \right) \|w_T\|^2. \end{aligned}$$

Thus, in view of (4.144) we test (4.140) with $\varphi_k = -\tau^{-1}\Delta^{-1}\xi_k$ in order to obtain

$$\begin{aligned}
 (4.146) \quad & \tau_{k,n}\|\xi_k\|_{I_n}^2 + \frac{\tau_{k,n}}{2} \left(\|\nabla\Delta^{-1}\xi_{k,n}\|^2 - \|\nabla\Delta^{-1}\xi_{k,n+1}\|^2 \right) - (\tau_{k,n}\Delta^{-1}\xi_k, \tilde{d}\xi_k)_{I_n} \\
 & \leq \frac{\tau_{k,n}}{2} \|\xi_k\|_{I_n}^2 + \frac{\tau_{k,n}}{2} \int_{I_n} (t_n - t)^2 \|\partial_t w(t)\|^2 dt \\
 & \quad - (\tau_{k,n}\Delta^{-1}\xi_k, \tilde{d}(\pi_k w - w))_{I_n}.
 \end{aligned}$$

As in the first part of the proof, the terms containing \tilde{d} are handled by Young's and Minkowski's inequality as well as the compactness of the Laplacian operator. In particular

$$\begin{aligned}
 & (\tau_{k,n}\Delta^{-1}\xi_k, \tilde{d}\xi_k)_{I_n} - (\tau_{k,n}\Delta^{-1}\xi_k, \tilde{d}(\pi_k w - w))_{I_n} = \tau_{k,n}(\Delta^{-1}\xi_k, \tilde{d}\varepsilon_k)_{I_n} \\
 & \leq \frac{1}{2} \left(\tau_{k,n}\|\Delta^{-1}\xi_k\|_{I_n}^2 + \|\tilde{d}\|_{\infty,\infty}^2 \|\varepsilon_k\|_{I_n}^2 \right) \\
 & \leq \frac{1}{2} \left(\tau_{k,n}C\|\xi_k\|_{I_n}^2 + \|\tilde{d}\|_{\infty,\infty}^2 \|\varepsilon_k\|_{I_n}^2 \right) \\
 & \leq \frac{C}{2} \left(\tau_{k,n}\|\varepsilon_k\|_{I_n}^2 + \tau_{k,n}\|\eta_k\|_{I_n}^2 + \|\varepsilon_k\|_{I_n}^2 \right) \\
 & \leq \frac{C}{2} \left(k_n^2 \tau_{k,n} \|\partial_t w\|_{I_n}^2 + \|\varepsilon_k\|_{I_n}^2 \right),
 \end{aligned}$$

using in the last step the approximation property (4.20) and Theorem 4.4.2 restricted to I_n .

While the second term in the expression above, thanks to the $L^2(I, L^2(\Omega))$ -norm estimate from Theorem 4.4.2, already displays what we are seeking, the first one requires some more steps. In particular, we observe that

$$\tau_{k,n} \leq T - t_n + \tilde{k}k_{n+1} \leq (1 + \tilde{k})(T - t_n) \leq (1 + \tilde{k})(T - t),$$

for $n = 1, \dots, N-1$, and $\tau_{k,N} = k_N$. Then, summing for $n = 1, \dots, N$, we have

$$\begin{aligned}
 (4.147) \quad & \sum_{n=1}^N k_n^2 \tau_{k,n} \|\partial_t w\|_{I_n}^2 = \sum_{n=1}^{N-1} k_n^2 \tau_{k,n} \int_{I_n} \|\partial_t w\|^2 dt + k_N^2 \tau_{k,N} \int_{I_N} \|\partial_t w\|^2 dt \\
 & \leq (1 + \kappa)k^2 \int_{I_n} (T - t) \|\partial_t w\|^2 dt + k^3 \int_{I_N} \|\partial_t w\|^2 dt \\
 & \leq Ck^2 \int_I (T - t) \|\partial_t w\|^2 dt.
 \end{aligned}$$

Now, summing also (4.146) for $n = 1, \dots, N$, we have

$$\begin{aligned}
 (4.148) \quad & T\|\nabla\Delta^{-1}\xi_{k,1}\|^2 + \sum_{n=1}^N \tau_{k,n}\|\xi_k\|_{I_n}^2 \leq \tilde{k}\|\nabla\Delta^{-1}\xi_k\|_I^2 \\
 & \quad + \sum_{n=1}^N \tau_{k,n} \int_{I_n} (t_n - t)^2 \|\partial_t w\|^2 dt + Ck^2 \int_I (T - t) \|\partial_t w\|^2 dt + \|\varepsilon_k\|_I^2,
 \end{aligned}$$

compare with [61, Equation (5.9)].

For the first term in the right-hand side, we use (4.143), the second follows as

in (4.93), while for the last two, we use (4.34b) and Theorem 4.4.2, respectively. All together this yields

$$(4.149) \quad T \|\nabla \Delta^{-1} \xi_{k,1}\|^2 + \sum_{n=1}^N \tau_{k,n} \|\xi_k\|_{I_n}^2 \leq Ck^2 \|w_T\|,$$

which in turn shows the claim. \square

With the last result at hand, the proof of Theorem 4.4.4 has been concluded and we can derive the error estimate for the error in the primal variable $e_k = u - u_k$. In the following, we use the state space U as defined in (3.16). The following result has been derived by the author of this thesis in [55, Theorem 4.2]

Theorem 4.4.9. *Let $u \in U$ and $u_k \in U_k$ be solutions of (2.28) and (2.29), respectively, with $f(t, x) = q(t)g(x) \in L^\infty(I, L^2(\Omega))$ and $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$. Then, for the error induced by the discretization in time there holds*

$$(4.150) \quad \|u - u_k\|_{L^\infty(I, L^2(\Omega))} \leq Ck \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \left(\|f\|_{L^\infty(I, L^2(\Omega))} + \|u_0\|_{\dot{H}^2(\Omega)} + \|d(\cdot, \cdot, 0)\|_{L^\infty(I \times \Omega)} \right).$$

Proof. In every time interval, we split the error into

$$\|e_k\|_{L^\infty(I_n, L^2(\Omega))} \leq \underbrace{\|u(\cdot) - u(t_n)\|_{L^\infty(I_n, L^2(\Omega))}}_{(a_1)} + \underbrace{\|u(t_n) - u_k(\cdot)\|_{L^\infty(I_n, L^2(\Omega))}}_{(a_2)}$$

and we analyze the two terms separately. Then, summing the estimates for any I_n , $n = 1, \dots, N$ we obtain the thesis. As in Theorem 4.3.6, with no loss of generality, we focus on the last time interval I_N .

(a_1) For a generic fixed time $\hat{t} \in I_N$, we start the derivation considering the interpolation error $u(\hat{t}) - u(t_N)$.

Consider the solutions w and \hat{w} of Problem 4.1.6(a) and (b), respectively, with terminal value w_T to be specified later. Integration by parts in time of (4.11) and (4.12) leads to

$$\begin{aligned} -(\varphi(T), w(T)) + (\varphi(0), w(0)) + (\partial_t \varphi, w)_I + (\nabla \varphi, \nabla w)_I + (\varphi, \tilde{d}w)_I &= 0, \\ -(\varphi(\hat{t}), \hat{w}(\hat{t})) + (\varphi(0), \hat{w}(0)) + (\partial_t \varphi, \hat{w})_{\hat{I}} + (\nabla \varphi, \nabla \hat{w})_{\hat{I}} + (\varphi, \tilde{d}\hat{w})_{\hat{I}} &= 0, \end{aligned}$$

for any $\varphi \in W(0, T)$.

Then, setting $\varphi = u$ and using the abbreviation $d(u)$, the state equation (2.28) yields

$$\begin{aligned} -(u(T), w(T)) + (u(0), w(0)) + (qg, w)_I - (d(u), w)_I + (u, \tilde{d}w)_I &= 0, \\ -(u(\hat{t}), \hat{w}(\hat{t})) + (u(0), \hat{w}(0)) + (qg, \hat{w})_{\hat{I}} - (d(u), \hat{w})_{\hat{I}} + (u, \tilde{d}\hat{w})_{\hat{I}} &= 0. \end{aligned}$$

Recalling that $w(T) = w(\hat{t}) = w_T$, subtracting the equalities above, we

have

$$\begin{aligned}
 (4.151) \quad (u(\hat{t}) - u(T), w_T) = & (u(0), \hat{w}(0) - w(0)) + (qg, \hat{w} - w)_{\hat{I}} - (qg, w)_{I \setminus \hat{I}} \\
 & + \underbrace{(u, \tilde{d}(\hat{w} - w))_{\hat{I}}}_{(b_1)} - \underbrace{(u, \tilde{d}w)_{I \setminus \hat{I}}}_{(b_2)} \\
 & + \underbrace{(d(u), w - \hat{w})_{\hat{I}}}_{(b_3)} + \underbrace{(d(u), w)_{I \setminus \hat{I}}}_{(b_4)},
 \end{aligned}$$

and we analyze separately the terms.

(b₁) The boundedness of $\|\tilde{d}\|_{L^\infty(I \times \Omega)} \leq c$ entails

$$(u, \tilde{d}(\hat{w} - w))_{\hat{I}} \leq c \|u\|_{L^\infty(I, L^2(\Omega))} \|\hat{w} - w\|_{L^1(I, L^2(\Omega))}.$$

(b₂) Exploiting again the boundedness of \tilde{d} in $L^\infty(I \times \Omega)$, noting that $|T - \hat{t}| \leq k$, we have

$$\begin{aligned}
 -(u, \tilde{d}w)_{I \setminus \hat{I}} & \leq \left| \int_{\hat{t}}^T (u, \tilde{d}w) dt \right| \\
 & \leq ck \|u\|_{L^\infty(I, L^2(\Omega))} \|w\|_{L^\infty(I, L^2(\Omega))}
 \end{aligned}$$

(b₃) The Lipschitz properties of $d(\cdot, \cdot, u)$ and the boundedness of $d(\cdot, \cdot, 0)$ in $L^\infty(\hat{I}, L^2(\Omega))$ yield

$$\begin{aligned}
 (d(u), w - \hat{w})_{\hat{I}} & = (d(u) - d(0), w - \hat{w})_{\hat{I}} + (d(0), w - \hat{w})_{\hat{I}} \\
 & \leq \|d(u) - d(0)\|_{L^\infty(\hat{I}, L^2(\Omega))} \|w - \hat{w}\|_{L^1(\hat{I}, L^2(\Omega))} \\
 & \quad + \|d(0)\|_{L^\infty(\hat{I}, L^2(\Omega))} \|w - \hat{w}\|_{L^1(\hat{I}, L^2(\Omega))} \\
 & \leq c(\|u\|_{L^\infty(\hat{I}, L^2(\Omega))} + \|d(0)\|_{L^\infty(\hat{I}, L^2(\Omega))}) \\
 & \quad \cdot \|w - \hat{w}\|_{L^1(\hat{I}, L^2(\Omega))}.
 \end{aligned}$$

(b₄) Using the same argument as before, we conclude

$$\begin{aligned}
 (d(u), w)_{I \setminus \hat{I}} & = (d(u) - d(0), w)_{I \setminus \hat{I}} + (d(0), w)_{I \setminus \hat{I}} \\
 & \leq ck(\|u\|_{L^\infty(I \times \Omega)} + \|d(0)\|_{L^\infty(I \times \Omega)}) \|w\|_{L^\infty(I, L^2(\Omega))}.
 \end{aligned}$$

Back to (4.151), we pick $w_T = u(\hat{t}) - u(T)$ and obtain

$$\begin{aligned}
 \|u(\hat{t}) - u(T)\|^2 & \leq c \left(\|w - \hat{w}\|_{L^1(\hat{I}, L^2(\Omega))} + \|(w - \hat{w})(0)\|_{H^{-2}(\Omega)} \right. \\
 & \quad \left. + k\|w\|_{L^\infty(I, L^2(\Omega))} \right) \cdot \left(\|qg\|_{L^\infty(I, L^2(\Omega))} + \|u_0\|_{\dot{H}^2(\Omega)} \right. \\
 & \quad \left. + \|d(0)\|_{L^\infty(I \times \Omega)} + \|u\|_{L^\infty(I, L^2(\Omega))} + \|u\|_{L^\infty(I \times \Omega)} \right).
 \end{aligned}$$

Using Propositions 4.2.4 and 3.2.2, together with Lemma 4.4.5, after dividing by $\|w_T\|$, we conclude

$$\begin{aligned}
 (4.152) \quad \|u(\hat{t}) - u(T)\| & \leq ck \log \left(\frac{T}{k} + 1 \right)^{\frac{1}{2}} \left(\|q\|_{L^\infty(I, \mathbb{R}^m)} \|g\|_{L^2(\Omega)} + \|u_0\|_{\dot{H}^2(\Omega)} \right. \\
 & \quad \left. + \|d(0)\|_{L^\infty(I \times \Omega)} \right).
 \end{aligned}$$

(a₂) To obtain the error inside the time interval I_N , we set

$$w_T = u(t_N) - u_{k,N} = u(T) - u_{k,N}$$

in (4.11) and in (4.13), the latter defined through \tilde{d} , and get

$$B(\varphi, w) + (\varphi, \tilde{d}w)_I = B(\varphi, w_k) + (\varphi, \tilde{d}w_k)_I = (\varphi_N, u(T) - u_{k,N}),$$

for $\varphi \in U_k$. In particular, testing the relation above with $\varphi = u - u_k$, observing that for $\varphi_k \in U_k$ it holds

$$(4.153) \quad B(u - u_k, \varphi_k) = -(d(u) - d(u_k), \varphi_k)_I = -((u - u_k)\tilde{d}, \varphi_k)_I,$$

and making use of (4.19), we have

$$\begin{aligned} \|u(T) - u_{k,N}\|^2 &= B(u - u_k, w) + (u - u_k, \tilde{d}w)_I \\ &= B(u - u_k, w - w_k) - (\tilde{d}(u - u_k), w_k)_I + (\tilde{d}(u - u_k), w)_I \\ &= B(u, w - w_k) + (u_k, \tilde{d}(w - w_k))_I + (\tilde{d}(u - u_k), w - w_k)_I \\ &= (qg, w - w_k)_I + (u_0, w(0) - w_k(0)) - \underbrace{(d(u), w - w_k)_I}_{(c_1)} \\ &\quad + \underbrace{(u_k, \tilde{d}(w - w_k))_I}_{(c_2)} + \underbrace{(\tilde{d}(u - u_k), w - w_k)_I}_{(c_3)}, \end{aligned}$$

using (2.28) in the last step.

We consider the three terms separately.

(c₁) Observing that $L^\infty(I, H_0^1(\Omega)) \hookrightarrow L^\infty(I, L^2(\Omega))$, the stability of u from Proposition 3.2.2, the Lipschitz continuity of $d(\cdot, \cdot, u)$ and the boundedness of $d(\cdot, \cdot, 0)$ in $L^\infty(I, L^2(\Omega))$, yield

$$\begin{aligned} -(d(u), w - w_k)_I &\leq \left(\|d(u) - d(0)\|_{L^\infty(I, L^2(\Omega))} + \|d(0)\|_{L^\infty(I, L^2(\Omega))} \right) \\ &\quad \cdot \|w - w_k\|_{L^1(I, L^2(\Omega))} \\ &\leq c \left(\|u\|_{L^\infty(I, L^2(\Omega))} + \|d(0)\|_{L^\infty(I, L^2(\Omega))} \right) \\ &\quad \cdot \|w - w_k\|_{L^1(I, L^2(\Omega))} \\ &\leq c \left(\|qg\|_I + \|u_0\|_V + \|d(0)\|_{L^\infty(I, L^2(\Omega))} \right) \\ &\quad \cdot \|w - w_k\|_{L^1(I, L^2(\Omega))}. \end{aligned}$$

(c₂) The boundedness of \tilde{d} in $L^\infty(I \times \Omega)$ and Proposition 3.2.3, yield

$$\begin{aligned} (u_k, \tilde{d}(w - w_k))_I &\leq \|u_k\|_{L^\infty(I, L^2(\Omega))} \|w - w_k\|_{L^1(I, L^2(\Omega))} \\ &\leq c \left(\|qg\|_I + \|u_0\|_V + \|d(0)\|_I \right) \|w - w_k\|_{L^1(I, L^2(\Omega))}. \end{aligned}$$

(c₃) Using the Lipschitz continuity of $d(\cdot, \cdot, u)$ combined with the definition and boundedness of \tilde{d} , we get

$$\begin{aligned} (\tilde{d}(u - u_k), w - w_k)_I &= (d(u) - d(u_k), w - w_k)_I \\ &= (d(u) - d(0), w - w_k)_I + (d(0) - d(u_k), w - w_k)_I \\ &\leq c \left(\|u\|_{L^\infty(I, L^2(\Omega))} + \|u_k\|_{L^\infty(I, L^2(\Omega))} \right) \|w - w_k\|_{L^1(I, L^2(\Omega))} \\ &\leq c \left(\|qg\|_I + \|u_0\|_{H_0^1(\Omega)} + \|d(0)\|_I \right) \|w - w_k\|_{L^1(I, L^2(\Omega))}, \end{aligned}$$

using in the last step the stability of the solutions u and u_k from Propositions 3.2.2 and Proposition 3.2.3, respectively.

Summing up, for the error inside the time interval we have

$$\begin{aligned} \|u(T) - u_{k,N}\|^2 &\leq c \left(\|w - w_k\|_{L^I(I, L^2(\Omega))} + \|w(0) - w_k(0)\|_{H^{-2}(\Omega)} \right) \\ &\quad \cdot \left(\|qg\|_{L^\infty(I, L^2(\Omega))} + \|u_0\|_{\dot{H}^2(\Omega)} + \|d(\cdot, \cdot, 0)\|_{L^\infty(I, L^2(\Omega))} \right). \end{aligned}$$

In conclusion, combining (4.152) with the inequality above, and thanks to Theorem 4.4.4, we obtain the thesis. \square

4.4.2 Spatial error

We split the error in the auxiliary variables into a projection and discretization error by means of the L^2 -projection in space. With a little abuse of notation, we denote the error in the auxiliary variables v_k and v_{kh} , solutions of Problem 4.1.8(a) and (b), respectively, as

$$\begin{aligned} \varepsilon_h &= v_k - v_{kh} \\ &= v_k - P_h v_k + P_h v_k - v_{kh} \\ &= \eta_h + \xi_h. \end{aligned}$$

Then, we define the following auxiliary problem

$$(4.154) \quad B(\varphi_k, \bar{w}_k) + (\varphi_k, \hat{d}\bar{w}_k)_I = (\varphi_k, \varepsilon_h)_I, \quad \forall \varphi_k \in U_k,$$

for which we have, compare with [68],

$$(4.155) \quad \|\bar{w}_k\|_I + \|\Delta \bar{w}_k\|_I \leq C \|\varepsilon_h\|_I.$$

Further, we denote the projection error for \bar{w}_k with $\bar{\eta}_h = \bar{w}_k - P_h \bar{w}_k$. To avoid confusion, we remark that in the following two results the discretization error ξ_h and the projection error η_h are those defined with respect to v_k

Lemma 4.4.10. *For the discretization error ξ_h and the projection errors $\eta_h, \bar{\eta}_h$ the following estimates hold*

$$(4.156a) \quad B(\eta_h, \varphi) = (\nabla \eta_h, \nabla \varphi)_I, \quad \forall \varphi \in U_{kh} \cup U_k$$

$$(4.156b) \quad B(\varphi, \bar{\eta}_h) = (\nabla \varphi, \nabla \bar{\eta}_h)_I, \quad \forall \varphi \in U_{kh} \cup U_k$$

$$(4.156c) \quad B(\eta_h, \bar{\eta}_h) \leq \|\nabla \eta_h\|_I \|\nabla \bar{\eta}_h\|_I + C \|\eta_h\|_I \|\varepsilon_h\|_I.$$

Proof. Relations (4.156a) and (4.156b) are shown in [62, Lemma 5.6] for $\varphi \in U_{kh}$. Considering as in our case polynomial piecewise constant in time, the claim holds for $\varphi \in U_k$ as well.

Equation (4.156c) follows as in [62, Lemma 5.8] \square

Lemma 4.4.11. *For the discretization error ξ_h and the projection error η_h there holds*

$$(4.157) \quad \|\nabla \xi_h\|_I \leq C (\|\varepsilon_h\|_I + \|\nabla \eta_h\|_I).$$

Proof. By means of the definition of $B(\cdot, \cdot)$ (2.15), relations (4.18) and (4.156a), we have

$$\begin{aligned}\|\nabla \xi_h\|_I^2 &= (\nabla \xi_h, \nabla \xi_h)_I \leq |B(\xi_h, \xi_h)| = |B(\varepsilon_h - \eta_h, \xi_h)| \\ &\leq (\hat{d}\varepsilon_h, \xi_h)_I + (\nabla \eta_h, \nabla \xi_h)_I \\ &\leq \left(\|\hat{d}\|_{L^\infty(I \times \Omega)} \|\varepsilon_h\|_I \|\xi_h\|_I + \|\nabla \eta_h\|_I \|\nabla \xi_h\|_I \right) \\ &\leq C \left(\|\varepsilon_h\|_I + \|\nabla \eta_h\|_I \right) \|\nabla \xi_h\|_I.\end{aligned}$$

The claim follows dividing by $\|\nabla \xi_h\|_I$. \square

Remark 4.4.12. Comparing the Lemma above with [62, Lemma 5.7], where the analogous is shown for the linear case, we observe that the price to pay for having the semi-linear term is the presence of $\|\varepsilon_h\|_I$. This is indeed what one expects because the Galerkin orthogonality in the linear case is substituted by the relation (4.18) when considering semi-linear equation.

Further, it also highlights how estimates in the $L^2(I, L^2(\Omega))$ -norm for the error in the auxiliary variables are again necessary.

Using a technique similar to [68, Theorem 4.2], we start with the error in the $L^2(I, L^2(\Omega))$ -norm between the solutions of Problems 4.1.8 (a) and (b).

Lemma 4.4.13. For the error between the solutions v_k and v_{kh} of (4.15) and 4.16, respectively, there holds

$$(4.158) \quad \|v_k - v_{kh}\|_I \leq Ch^2 \|\nabla v_0\|.$$

Proof. Testing (4.154) with $\varphi_k = \xi_h$, we have

$$\begin{aligned}(\xi_h, \varepsilon_h)_I &= B(\xi_h, \bar{w}_k) + (\xi_h, \hat{d}\bar{w}_k)_I \\ &= B(\varepsilon_h - \eta_h, \bar{w}_k) + (\varepsilon_h - \eta_h, \hat{d}\bar{w}_k)_I \\ &= -(\varepsilon_h \hat{d}, \bar{w}_k)_I - (\nabla \eta_h, \nabla \bar{w}_k)_I + (\varepsilon_h - \eta_h, \hat{d}\bar{w}_k)_I \\ &= -(\nabla \eta_h, \nabla \bar{w}_k)_I - (\eta_h, \hat{d}\bar{w}_k)_I,\end{aligned}$$

using (4.18) and (4.156a). This implies

$$\begin{aligned}\|\varepsilon_h\|_I^2 &= (\eta_h, \varepsilon_h)_I - (\nabla \eta_h, \nabla \bar{w}_k)_I - (\eta_h, \hat{d}\bar{w}_k)_I \\ &= (\eta_h, \varepsilon_h)_I + (\eta_h, \Delta \bar{w}_k)_I - (\eta_h, \hat{d}\bar{w}_k)_I \\ &\leq \|\eta_h\|_I \left(\|\varepsilon_h\|_I + \|\Delta \bar{w}_k\|_I + \|\hat{d}\|_{L^\infty(I \times \Omega)} \|\bar{w}_k\|_I \right) \\ &\leq C \|\eta_h\|_I \|\varepsilon_h\|_I,\end{aligned}$$

using (4.155). Then, the well-known estimate

$$\|\eta_h\|_I \leq Ch^2 \|\nabla^2 v_k\|_I,$$

and the fact that, being the domain polygonal and convex, we have,

$$\|\nabla^2 v_k\|_I \leq C \|\Delta v_k\|_I,$$

together with the stability of v_k , yield

$$(4.159) \quad \|\varepsilon_h\| \leq Ch^2 \|\Delta v_k\|_I \leq Ch^2 \|\nabla v_0\|.$$

\square

Next two results have to be compared with [61, Lemma 5.6(b)-Lemma 5.7(b)] where they are shown in the linear setting. With respect to their setting, where the claim is entailed by the Galerkin orthogonality, in the non-linear setting, we have to exploit again the auxiliary problem (4.154) defined above.

Lemma 4.4.14. *For the error between the solutions v_k and v_{kh} of (4.15) and (4.16), respectively, there holds*

$$(4.160) \quad \|v_k - v_{kh}\|_I \leq C\sqrt{T}h^2\|\Delta v_0\|.$$

Proof. We test (4.154) with $\varphi_k = \varepsilon_h$ and, recalling $\bar{\eta}_h = \bar{w}_k - P_h\bar{w}_k$, thanks to (4.18), we obtain

$$(4.161) \quad \begin{aligned} \|\varepsilon_h\|_I^2 &= B(\varepsilon_h, \bar{w}_k) + (\varepsilon_h, \hat{d}\bar{w}_k)_I \\ &= B(\varepsilon_h, \bar{\eta}_h) + B(\varepsilon_h, P_h\bar{w}_k) + (\varepsilon_h, \hat{d}\bar{\eta}_h)_I + (\varepsilon_h, \tilde{d}P_h\bar{w}_k)_I \\ &= B(\varepsilon_h, \bar{\eta}_h) - (\hat{d}\varepsilon_h, P_h\bar{w}_k)_I + (\varepsilon_h, \hat{d}\bar{\eta}_h)_I + (\varepsilon_h, \tilde{d}P_h\bar{w}_k)_I \\ &= B(\eta_h + \xi_h, \bar{\eta}_h) + (\varepsilon_h, \hat{d}\bar{\eta}_h)_I \\ &\leq \|\nabla\eta_h\|_I\|\nabla\bar{\eta}_h\|_I + C\|\eta_h\|_I\|\varepsilon_h\|_I + \|\nabla\xi_h\|_I\|\nabla\bar{\eta}_h\|_I \\ &\quad + \|\hat{d}\|_{L^\infty(I\times\Omega)}\|\varepsilon_h\|_I\|\bar{\eta}_h\|_I \end{aligned}$$

using in the last step (4.156c), (4.156b) and the boundedness of \hat{d} , respectively. For the L^2 -projection error there holds the following estimates

$$(4.162) \quad \|\phi - P_h\phi\|_I \leq Ch^2\|\nabla^2\phi\|_I, \quad \|\nabla(\phi - P_h\phi)\|_I \leq Ch\|\nabla^2\phi\|_I$$

which, together with (4.157), lead to

$$(4.163) \quad \begin{aligned} \|\varepsilon_h\|_I^2 &\leq C \left(\underbrace{h^2\|\nabla^2 v_k\|_I\|\nabla^2 \bar{w}_k\|_I + h^2\|\nabla^2 v_k\|_I\|\varepsilon_h\|_I}_{(a_1)} \right. \\ &\quad \left. + \underbrace{h\|\nabla^2 \bar{w}_k\|_I(\|\varepsilon_h\|_I + \|\nabla\eta_h\|_I)}_{(a_2)} \right. \\ &\quad \left. + \underbrace{h^2\|\hat{d}\|_{L^\infty(I\times\Omega)}\|\nabla^2 \bar{w}_k\|_I\|\varepsilon_h\|_I}_{(a_3)} \right). \end{aligned}$$

We consider the three parts separately.

(a_1) The domain Ω being polygonal and convex, (4.155) and Proposition 4.2.9 lead to

$$(4.164) \quad \begin{aligned} (a_1) &\leq Ch^2\|\Delta v_k\|_I(\|\Delta \bar{w}_k\|_I + \|\varepsilon_h\|_I) \\ &\leq Ch^2\|\Delta v_k\|_I\|\varepsilon_h\|_I \\ &\leq Ch^2\sqrt{T} \max_{n=1,\dots,N} \|\Delta v_{k,n}\| \|\varepsilon_h\|_I \\ &\leq C\sqrt{T}h^2\|\Delta v_0\| \|\varepsilon_h\|_I. \end{aligned}$$

(a_2) In a first step, we use Lemma 4.4.13 and (4.162) to bound the terms inside the inner brackets. Then, we conclude thanks to (4.155) and Proposi-

tion 4.2.9

$$\begin{aligned}
 (4.165) \quad (a_2) &\leq Ch\|\Delta\bar{w}_k\|\left(h^2\|\nabla v_0\| + h\|\Delta v_k\|_I\right) \\
 &\leq C(h^2\|\Delta v_k\|_I + h^3\|\nabla v_0\|)\|\varepsilon_h\|_I \\
 &\leq C(\sqrt{T}h^2\|\Delta v_0\|_I + h^3\|\nabla v_0\|)\|\varepsilon_h\|_I.
 \end{aligned}$$

(a_3) Lemma 4.4.13 and (4.155) give

$$\begin{aligned}
 (4.166) \quad (a_3) &\leq Ch^4\|\Delta\bar{w}_k\|_I\|\nabla u_0\| \\
 &\leq Ch^4\|\nabla u_0\|\|\varepsilon_h\|_I.
 \end{aligned}$$

In conclusion, combining (4.164) with (4.165) and (4.166) in (4.163), and considering the leading term h^2 , we obtain, after division for $\|\varepsilon_h\|_I$,

$$\|\varepsilon_h\|_I \leq C\sqrt{T}h^2\|\Delta v_0\|.$$

□

Lemma 4.4.15. *For the error at the final time between the solutions v_k and v_{kh} of (4.15) and (4.16), respectively, there holds*

$$(4.167) \quad \|v_{k,N} - v_{kh,N}\|_I \leq Ch^2\|\Delta v_0\|.$$

Proof. We introduce the Ritz projection $R_h: H_0^1(\Omega) \rightarrow V_h$

$$(\nabla R_h v, \varphi_h) = (\nabla v, \nabla \varphi_h), \quad \forall \varphi_h \in V_h,$$

with the corresponding estimate

$$(4.168) \quad \|R_h v - v\|_I \leq Ch^2\|\Delta v\|_I,$$

see [89]. We split the error in the auxiliary variable as

$$\begin{aligned}
 \varepsilon_h &= v_k - v_{kh} \\
 &= v_k - R_h v_k + R_h v_k - v_{kh} \\
 &= \eta_h + R_h \varepsilon_h
 \end{aligned}$$

noting that $R_h v_{kh} = v_{kh}$.

We observe that, by means of (4.18), for any $\varphi_{kh} \in U_{kh}$, it holds

$$\begin{aligned}
 B(R_h \varepsilon, \varphi_{kh}) &= B(\varepsilon_h, \varphi_{kh}) - B(\eta_h, \varphi_{kh}) \\
 &= -(\hat{d}\varepsilon_h, \varphi_{kh})_I - B(\eta_h, \varphi_{kh}),
 \end{aligned}$$

which, restricted on the time interval I_n , reads

$$\begin{aligned}
 (4.169) \quad (\nabla R_h \varepsilon_h, \nabla \varphi_{kh})_{I_n} + ([R_h \varepsilon_h]_{n-1}, \varphi_{kh,n}) &= -(\hat{d}\varepsilon_h, \varphi_{kh})_{I_n} \\
 &\quad - (\nabla \eta_h, \nabla \varphi_{kh})_{I_n} - ([R_h \eta_h]_{n-1}, \varphi_{kh,n}),
 \end{aligned}$$

for any $\varphi_{kh} \in \mathcal{P}_0(I_n, V_h)$. We test (4.169) with $\varphi_{kh} = t_{n-1}R_h \varepsilon_h$ and get

$$\begin{aligned}
 (4.170) \quad t_{n-1}(\nabla R_h \varepsilon_h, \nabla R_h \varepsilon_h)_{I_n} + t_{n-1}([R_h \varepsilon_h]_{n-1}, R_h \varepsilon_{h,n}) &= -t_{n-1}(\hat{d}\varepsilon_h, R_h \varepsilon_h)_{I_n} \\
 &\quad - t_{n-1}(\nabla \eta_h, \nabla R_h \varepsilon_h)_{I_n} - t_{n-1}([R_h \eta_h]_{n-1}, R_h \varepsilon_{h,n}),
 \end{aligned}$$

We move our attention to the first term in the right-hand side containing \hat{d} . Using Young's inequality and the boundedness of \hat{d} , we have

$$\begin{aligned}
 (4.171) \quad -t_{n-1}(\hat{d}\varepsilon_h, R_h\varepsilon_h)_{I_n} &= -t_{n-1}(\hat{d}(\eta_h + R_h\varepsilon_h), R_h\varepsilon_h)_{I_n} \\
 &\leq \frac{t_{n-1}^2}{2}\|\eta_h\|_{I_n}^2 + \frac{C}{2}\|R_h\varepsilon_h\|_{I_n}^2 - t_{n-1}\|\sqrt{\hat{d}}R_h\varepsilon_h\|_{I_n}^2 \\
 &= \frac{t_{n-1}^2}{2}\|\eta_h\|_{I_n}^2 + \frac{C}{2}k_n\|R_h\varepsilon_{h,n}\|^2 - t_{n-1}\|\sqrt{\hat{d}}R_h\varepsilon_h\|_{I_n}^2,
 \end{aligned}$$

where in the last step we used

$$\|R_h\varepsilon_h\|_{I_n}^2 = \int_{I_n} \|R_h\varepsilon_h\|^2 dt = k_n\|R_h\varepsilon_{h,n}\|^2.$$

The terms left in (4.170) are handled as in [61, Lemma 5.7b] and therefore we omit the details. Inserting (4.171) in (4.170), we have

$$\begin{aligned}
 (4.172) \quad &t_n\|R_h\varepsilon_{h,n}\|^2 + 2t_{n-1}\|\nabla R_h\varepsilon_k\|_{I_n}^2 + t_{n-1}\|\sqrt{\hat{d}}R_h\varepsilon_h\|_{I_n}^2 \leq \\
 &t_{n-1}\|R_h\varepsilon_{h,n-1}\|^2 + k_n\|R_h\varepsilon_{h,n}\|^2 + \frac{t_{n-1}^2}{k_n}\|[\eta_h]_{n-1}\|^2 \\
 &+ k_n\|R_h\varepsilon_{h,n}\|^2 + t_{n-1}^2\|\eta_h\|_{I_n}^2 + Ck_n\|R_h\varepsilon_{h,n}\|^2.
 \end{aligned}$$

The second and third term in the left-hand side are positive and we erase them. Then, summing for $n = 2, \dots, N$ and observing that $k_1 = t_1$, we have

$$\begin{aligned}
 (4.173) \quad T\|R_h\varepsilon_{h,N}\|^2 &\leq k_1\|R_h\varepsilon_{h,1}\|^2 + (2+C)\sum_{n=2}^N k_n\|R_h\varepsilon_{h,n}\|^2 \\
 &+ \sum_{n=2}^N \frac{t_{n-1}^2}{k_n}\|[\eta_h]_{n-1}\|^2 + \sum_{n=2}^N t_{n-1}^2\|[\eta_h]_{n-1}\|^2 \\
 &\leq C\|R_h\varepsilon_h\|_I^2 + \sum_{n=2}^N \frac{t_{n-1}^2}{k_n}\|[\eta_h]_{n-1}\|^2 + \sum_{n=2}^N t_{n-1}^2\|[\eta_h]_{n-1}\|^2.
 \end{aligned}$$

Employing the splitting for ε_h we observe

$$\begin{aligned}
 \|R_h\varepsilon_h\|_I^2 &= \|\varepsilon_h\|_I^2 + \|\eta_h\|_I^2 - 2(\varepsilon_h, \eta_h)_I \\
 &\leq 2\|\varepsilon_h\|_I^2 + 2\|\eta_h\|_I^2
 \end{aligned}$$

thanks to Young's inequality. Then, back to (4.173) and exploiting again the

splitting for ε_h in the left-hand side, we obtain

$$\begin{aligned}
 T\|\varepsilon_{h,N}\|^2 &\leq C\left(T\|\eta_{h,N}\|^2 + \|\eta_h\|_I^2 + \sum_{n=2}^N \frac{t_{n-1}^2}{k_n} \|[\eta_h]_{n-1}\|^2 \right. \\
 &\quad \left. + \sum_{n=2}^N t_{n-1}^2 \|[\eta_h]_{n-1}\|^2\right) + C\|\varepsilon_h\|_I^2 \\
 (4.174) \quad &\leq CTh^4\left(\|\Delta v_{k,N}\|^2 + \|\Delta v_k\|_I^2 + \sum_{n=2}^N \frac{t_{n-1}}{k_n} \|[\Delta v_k]_{n-1}\|^2 \right. \\
 &\quad \left. + \sum_{n=2}^N t_{n-1} \|[\Delta v_h]_{n-1}\|_{I_n}^2\right) + CTh^4\|\Delta v_0\|_I^2 \\
 &\leq CTh^4\|\Delta v_0\|_I^2
 \end{aligned}$$

using Lemma 4.4.14 and, in the last step, Proposition 4.2.9. Division for T leads to the thesis

$$\|\varepsilon_{h,N}\| = \|v_{k,N} - v_{kh,N}\| \leq Ch^2\|\Delta v_0\|.$$

□

In view of the duality argument, we derive the error in the auxiliary dual variables solutions of Problems 4.1.7 based on the error estimates obtained above.

Lemma 4.4.16. *Let w_k and w_{kh} be solutions of (4.13) and (4.14), respectively. Then, it holds*

$$(4.175) \quad \|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} \leq ch^2\|w_T\|.$$

Proof. The definition of the H^{-2} -norm, which reads,

$$\|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} = \sup_{\psi \in \dot{H}^2(\Omega)} \frac{(w_{k,1} - w_{kh,1}, \psi)}{\|\psi\|_{H^2(\Omega)}}$$

suggests to bound the numerator with a quantity depending on $\|\Delta\psi\|$ and $\|w_T\|$. To achieve this, we proceed similarly to Lemma 4.3.10 selecting conveniently the test functions in the auxiliary problems.

We fix $\psi \in \dot{H}^2(\Omega)$ and set the initial data $v_0 = \psi$ in (4.15) and (4.16). Then we pick $\varphi_k = v_k$ in (4.13), $\varphi_{kh} = v_{kh}$ in (4.14), $\varphi_k = w_k$ in (4.15) and $\varphi_{kh} = w_{kh}$ in (4.16), obtaining

$$(4.176a) \quad B(v_k, w_k) + (\hat{d}v_k, w_k)_I = (\psi, w_{k,1}),$$

$$(4.176b) \quad B(v_k, w_k) + (v_k, \hat{d}w_k)_I = (v_{k,N}, w_T),$$

$$(4.176c) \quad B(v_{kh}, w_{kh}) + (\hat{d}v_{kh}, w_{kh})_I = (\psi, w_{kh,1}),$$

$$(4.176d) \quad B(v_{kh}, w_{kh}) + (v_{kh}, \hat{d}w_{kh})_I = (v_{kh,N}, w_T).$$

Then, we subtract (4.176c) to (4.176a), and using (4.17) and (4.18). we have

$$\begin{aligned}
(\psi, w_{k,1} - w_{kh}) &= B(v_k, w_k) - B(v_{kh}, w_{kh}) + (\hat{d}v_k, w_k)_I - (\hat{d}v_{kh}, w_{kh})_I \\
&= B(v_k - v_{kh}, w_k) + B(v_{kh}, w_k - w_{kh}) \\
&\quad + (\hat{d}v_k, w_k)_I - (\hat{d}v_{kh}, w_{kh})_I \\
&= B(v_k - v_{kh}, w_k - w_{kh}) - (\hat{d}(v_k - v_{kh}), w_{kh})_I \\
&\quad - (v_{kh}, \hat{d}(w_k - w_{kh}))_I + (\hat{d}v_k, w_k)_I - (\hat{d}v_{kh}, w_{kh})_I \\
&= B(v_k - v_{kh}, w_k - w_{kh}) - (\hat{d}v_k, w_{kh})_I \\
&\quad - (v_{kh}, \hat{d}w_k)_I + (v_{kh}, \hat{d}w_{kh})_I + (\hat{d}v_k, w_k)_I \\
&= B(v_k - v_{kh}, w_k - w_{kh}) + (v_k - v_{kh}, \hat{d}w_k)_I \\
&\quad - (v_k - v_{kh}, \hat{d}w_{kh})_I \\
&= B(v_k - v_{kh}, w_k - w_{kh}) + (v_k - v_{kh}, \hat{d}(w_k - w_{kh}))_I \\
&= (v_{k,N} - v_{kh,N}, w_T)
\end{aligned}$$

After this somehow lengthy computation, we conclude

$$\begin{aligned}
(\psi, w_{k,1} - w_{kh}) &= (v_{k,N} - v_{kh,N}, w_T) \\
&\leq \|v_{k,N} - v_{kh,N}\| \|w_T\| \\
&\leq Ch^2 \|\Delta v_0\| \|w_T\|,
\end{aligned}$$

thanks to Lemma 4.4.15. Then, the claim follows from the definition of the H^{-2} -norm recalling that $v_0 = \psi$. \square

Last result concerns the error in $L^1(I, L^2(\Omega))$ between the solutions of Problems 4.1.7. In the linear setting of Section 4.3.2, we have exploited the fact that the functions involved are piecewise constant in time. This in particular has lead to estimates in time in the L^2 -norm for the dual auxiliary variables. In the semi-linear setting, we will use a different approach, inspired by [61, Lemma 5.9], based on a L^1 -norm estimate in time.

Lemma 4.4.17. *Let w_k and w_{kh} be solutions of (4.13) and (4.14), respectively. Then there holds*

$$(4.177) \quad \|w_k - w_{kh}\|_{L^1(I, L^2(\Omega))} \leq Ch^2 \left(\log \frac{T}{k} + 1 \right)$$

Proof. Assuming we have already derived the estimate,

$$(4.178) \quad T \|w_{k,1} - w_{kh,1}\| \leq Ch^2 \|w_T\|,$$

the claim follows by

$$\begin{aligned}
\|w_k - w_{kh}\|_{L^1(I, L^2(\Omega))} &\leq \sum_{n=1}^N k_n \tau_{k,n}^{-1} \max_{n=1, \dots, N} (\tau_{k,n} \|w_{k,n} - w_{kh,n}\|) \\
&\leq Ch^2 \left(\log \frac{T}{k} + 1 \right) \|w_T\|.
\end{aligned}$$

Therefore, we focus on (4.178). To this end, we follow the bootstrap argument depicted in [61, Lemma 5.9] and set $v_0 = w_{k,1} - w_{kh,1}$ in (4.15) and (4.16).

Further, for a $\tilde{n} \leq N$, we pick as test functions $\varphi_k = w_k$ in (4.15), $\varphi_k = v_k$ in (4.13), $\varphi_{kh} = w_{kh}$ in (4.16) and $\varphi_{kh} = v_{kh}$ in (4.14), on $\{0\} \cup_{i=1}^{\tilde{n}} I_i$, and zero otherwise. This choice yields

$$\begin{aligned} (w_{k,1} - w_{kh,1}, w_{k,1}) &= B(v_k, w_k) + (\hat{d}v_k, w_k)_I = (v_{k,\tilde{n}}, w_{k,\tilde{n}+1}) \\ (w_{k,1} - w_{kh,1}, w_{kh,1}) &= B(v_{kh}, w_{kh}) + (\hat{d}v_{kh}, w_{kh})_I = (v_{kh,\tilde{n}}, w_{kh,\tilde{n}+1}), \end{aligned}$$

and, subtracting the equations above, we have

$$\begin{aligned} (4.179) \quad (w_{k,1} - w_{kh,1}, w_{k,1} - w_{kh,1}) &= B(v_k, w_k) - B(v_{kh}, w_{kh}) \\ &\quad + (\hat{d}v_k, w_k)_I - (\hat{d}v_{kh}, w_{kh})_I \\ &= (v_{k,\tilde{m}}, w_{k,\tilde{m}+1}) - (v_{kh,\tilde{m}}, w_{kh,\tilde{m}+1}) \end{aligned}$$

We focus now on the central part of the expression to obtain the differential equation connecting the left-hand side with the right hand-side. Using relations (4.17), (4.18), we have

$$\begin{aligned} B(v_k, w_k) - B(v_{kh}, w_{kh}) + \underbrace{(\hat{d}v_k, w_k)_I - (\hat{d}v_{kh}, w_{kh})_I}_{(a_1)} &= \\ &= B(v_k - v_{kh}, w_k) + B(v_{kh}, w_k - w_{kh}) + (a_1) \\ &= B(v_k - v_{kh}, w_k - w_{kh}) + B(v_k - v_{kh}, w_{kh}) + B(v_{kh}, w_k - w_{kh}) + (a_1) \\ &= B(v_k - v_{kh}, w_k - w_{kh}) - (\hat{d}(v_k - v_{kh}), w_{kh})_I - (v_{kh}, \hat{d}(w_k - w_{kh}))_I + (a_1) \\ &= \underbrace{B(v_k - v_{kh}, w_k - w_{kh}) + (\hat{d}(v_k - v_{kh}), w_k - w_{kh})_I}_{(a_2)} \\ &\quad - (\hat{d}(v_k - v_{kh}), w_k)_I - (v_{kh}, \hat{d}(w_k - w_{kh}))_I + (a_1) \\ &= (a_2) - (\hat{d}v_k, w_k)_I + (\hat{d}v_{kh}, w_k)_I - (v_{kh}, \hat{d}w_k) + (v_{kh}, \hat{d}w_{kh}) \\ &\quad + (\hat{d}v_k, w_k)_I - (\hat{d}v_{kh}, w_{kh})_I \\ &= B(v_k - v_{kh}, w_k - w_{kh}) + (\hat{d}(v_k - v_{kh}), w_k - w_{kh})_I. \end{aligned}$$

Going back to (4.179), we have the desired relation

$$\begin{aligned} (w_{k,1} - w_{kh,1}, w_{k,1} - w_{kh,1}) &= B(v_k - v_{kh}, w_k - w_{kh}) + (\tilde{d}(v_k - v_{kh}), w_k - w_{kh})_I \\ &= (v_{k,\tilde{m}}, w_{k,\tilde{m}+1}) - (v_{kh,\tilde{m}}, w_{kh,\tilde{m}+1}). \end{aligned}$$

From now on the proof is identical to [61, Lemma 5.9] and therefore we omit it. The arguments used are Lemmas 4.4.15 and (4.4.16) together with the stability of the auxiliary variables in Proposition 4.2.5. \square

We have now all the ingredients to derive the error induced by the space discretization. This results corresponds to [55, Theorem 4.4].

Theorem 4.4.18. *Let $u_k \in U_k$ and $u_{kh} \in U_{kh}$ be solutions of (2.29) and (2.30), respectively, with $f(t, x) = q(t)g(x) \in L^\infty(I, L^2(\Omega))$ and $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$. Then, for the error induced by the discretization in space there holds*

$$\begin{aligned} (4.180) \quad \|u_k - u_{kh}\|_{L^\infty(I, L^2(\Omega))} &\leq ch^2 \left(\log \frac{T}{k} + 1 \right) \left(\|qg\|_{L^\infty(I, L^2(\Omega))} + \|u_0\|_{H^2(\Omega)} \right. \\ &\quad \left. + \|d(\cdot, \cdot, 0)\|_{L^\infty(I \times \Omega)} \right). \end{aligned}$$

Proof. We can show the estimate on a single time interval I_n due to u_k and u_{kh} being constant on any time interval I_n , and with no loss of generality we consider the last time interval only. For a generic time interval I_n , we consider (4.13) and (4.14) on $I = (0, t_n)$ and, noting that $0 \leq \log(t_n/k) \leq \log(T/k)$, the proof follows similarly.

We set $w_T = u_{k,N} - u_{kh,N}$ in (4.13) and (4.14). Then, using (4.17) and noting that, for $\varphi_{kh} \in U_{kh}$, it holds

$$B(u_k - u_{kh}, \varphi_{kh}) = -(d(\cdot, \cdot, u_k) - d(\cdot, \cdot, u_{kh}), \varphi_{kh})_I = -((u_k - u_{kh})\hat{d}, \varphi_{kh})_I,$$

we have

$$\begin{aligned} \|u_{k,N} - u_{kh,N}\|^2 &= B(u_k - u_{kh}, w_k) + (u_k - u_{kh}, \hat{d}w_k)_I \\ &= B(u_k - u_{kh}, w_k - w_{kh}) - (\hat{d}(u_k - u_{kh}), w_{kh})_I \\ &\quad + (\hat{d}(u_k - u_{kh}), w_k)_I \\ &= B(u_k, w_k - w_{kh}) + (u_{kh}, \hat{d}(w_k - w_{kh}))_I \\ &\quad + (\hat{d}(u_k - u_{kh}), w_k - w_{kh})_I \\ &= (qg, w_k - w_{kh})_I + (u_0, w_{k,1} - w_{kh,1}) - \underbrace{(d(u_k), w_k - w_{kh})_I}_{(a_1)} \\ &\quad + \underbrace{(u_{kh}, \hat{d}(w_k - w_{kh}))_I}_{(a_2)} + \underbrace{(\hat{d}(u_k - u_{kh}), w_k - w_{kh})_I}_{(a_3)}, \end{aligned}$$

using (2.29) in the last step.. We analyze the three terms separately.

(a₁) The Lipschitz continuity of $d(\cdot, \cdot, u)$ and the boundedness of $d(\cdot, \cdot, 0)$ in $L^\infty(I, L^2(\Omega))$, give

$$\begin{aligned} -(d(u_k), w_k - w_{kh})_I &\leq c \left(\|d(u_k) - d(0)\|_{L^\infty(I, L^2(\Omega))} + \|d(0)\|_{L^\infty(I, L^2(\Omega))} \right) \\ &\quad \cdot \|w_k - w_{kh}\|_{L^1(I, L^2(\Omega))} \\ &\leq c \left(\|u_k\|_{L^\infty(I, L^2(\Omega))} + \|d(0)\|_{L^\infty(I, L^2(\Omega))} \right) \\ &\quad \cdot \|w_k - w_{kh}\|_{L^1(I, L^2(\Omega))}. \end{aligned}$$

(a₂) Recalling that \hat{d} is bounded, we have

$$(u_{kh}, \hat{d}(w_k - w_{kh}))_I \leq c \|u_{kh}\|_{L^\infty(I, L^2(\Omega))} \|w_k - w_{kh}\|_{L^1(I, L^2(\Omega))}$$

(a₃) For the last term, we rely again on the Lipschitz continuity of $d(\cdot, \cdot, u)$ to conclude

$$\begin{aligned} (\hat{d}(u_k - u_{kh}), w_k - w_{kh})_I &= (d(u_k) - d(u_{kh}), w_k - w_{kh})_I \\ &\leq c \left(\|u_k\|_{L^\infty(I, L^2(\Omega))} + \|u_{kh}\|_{L^\infty(I, L^2(\Omega))} \right) \\ &\quad \|w_k - w_{kh}\|_{L^1(I, L^2(\Omega))}. \end{aligned}$$

We now combine the previous inequalities and, thanks to the regularity of u_k and u_{kh} we obtain

$$\begin{aligned} \|u_{k,N} - u_{kh,N}\|^2 &\leq c \left(\|w_k - w_{kh}\|_{L^1(I, L^2(\Omega))} + \|w_{k,1} - w_{kh,1}\|_{H^{-2}(\Omega)} \right) \\ &\quad \cdot \left(\|qg\|_{L^\infty(I, L^2(\Omega))} + \|u_0\|_{H^2(\Omega)} + \|d(0)\|_{L^\infty(I \times \Omega)} \right). \end{aligned}$$

Then, using Lemma 4.4.16 and 4.4.17, and after division by $w_T = u_{k,N} - u_{kh,N}$, the claim follows. \square

Ultimately, combining the error estimate of Theorem 4.4.9 with the one given above, we recover the total discretization error stated in (4.2).

4.4.3 Extension to first-order integral constraints

In this section, we give an insight on how to extend the $L^\infty(I, H_0^1(\Omega))$ -norm estimates obtained in Section 4.3 to the case of a semi-linear state equation, as the one considered in (2.27b).

The approach used in Theorems 4.4.9 and 4.4.18 is based, among the others, on the Lipschitz continuity of $d(\cdot, \cdot, u)$ in $L^\infty(I, L^2(\Omega))$, [2, Section 9].

If we mimic the same approach, we would end up with estimates of the type

$$(d(u), w - w_k)_I \leq (\|d(u_k) - d(0)\|_{L^\infty(I, H_0^1(\Omega))} + \|d(0)\|_{L^\infty(I, H_0^1(\Omega))}) \cdot \|w - w_k\|_{L^\infty(I, H^{-1}(\Omega))}.$$

However, since $d(\cdot, \cdot, u)$ is not Lipschitz continuous in $L^\infty(I, H_0^1(\Omega))$, the inequality above does not hold. The growth conditions to impose on $d(\cdot, \cdot, \cdot)$ to have such property would lead to a degeneration of the semi-linear term into a linear term.

To overcome this issue and still keep the duality argument, one might use a different strategy and estimate the term above as

$$(d(u), w - w_k)_I \leq \|\nabla d(u)\|_{L^\infty(I, L^2(\Omega))} \|w - w_k\|_{L^\infty(I, H^{-1}(\Omega))}.$$

This would lead to an estimate of the type

$$(4.181) \quad \|\nabla d(u)\|_{L^\infty(I, L^2(\Omega))} \leq c \|d'(u)\|_{L^\infty(I \times \Omega)} \|\nabla u\|_{L^\infty(I, L^2(\Omega))}.$$

Then, the second term in the right-hand side is directly bounded by virtue of Proposition 3.2.2. Further, the boundedness in $L^\infty(I \times \Omega)$ of u , together with Assumption 2.3.1 for $d(t, x, u)$, should be enough to guarantee the boundedness of the remaining term. Then, once one has shown this last point, the derivation of the a priori error estimate should follow as in Section 4.3. What just claimed, needs to be performed also for the semi-discrete and discrete settings, using, in this case, the stability estimates from Propositions 3.2.3 and 3.2.4.

Clearly, the stability estimates for the discretization of Section 4.2.1 need to be extended as well. This can be done following the arguments of Section 4.2.2, that is, introducing same kind of linearizations \tilde{d} and \hat{d} as in (4.9) and (4.10), respectively.

Ultimately, what we expect is the same convergence order of the linear setting, that is, $\mathcal{O}(k + h)$.

4.5 Numerical results

After having derived theoretically the a priori error estimates in the previous sections, we now validate our findings numerically.

To this end, we consider the following linear differential equation

$$(4.182) \quad \begin{aligned} \partial_t u(t, x, y) - \Delta u(t, x, y) &= q(t)g(x, y) && \text{in } (0, T) \times \Omega, \\ u(0, \cdot) &= u_0 && \text{in } \Omega, \\ u(t, x, y) &= 0 && \text{on } (0, T) \times \partial\Omega, \end{aligned}$$

with $I = (0, 1)$ and $\Omega = (0, \pi)^2$. The following data are chosen

$$\begin{aligned} q(t) &= (1 - 2t), \\ g(x, y) &= \sin(x) \sin(y), \\ u_0 &= \sin(x) \sin(y), \end{aligned}$$

with the known solution given by

$$u(t, x, y) = (1 - t) \sin(x) \sin(y).$$

To solve the problem numerically, we use the DOpElib library [34], based on the finite element toolkits [4], which uses the Newton's method in order to solve the PDE. The time stepping scheme used is the implicit Euler scheme. As in the theoretical findings, we analyze separately the influence of the time and spatial discretization. Since the plots of the errors are indistinguishable for different level of refinements, the pictures below will exhibit only one level of refinement.

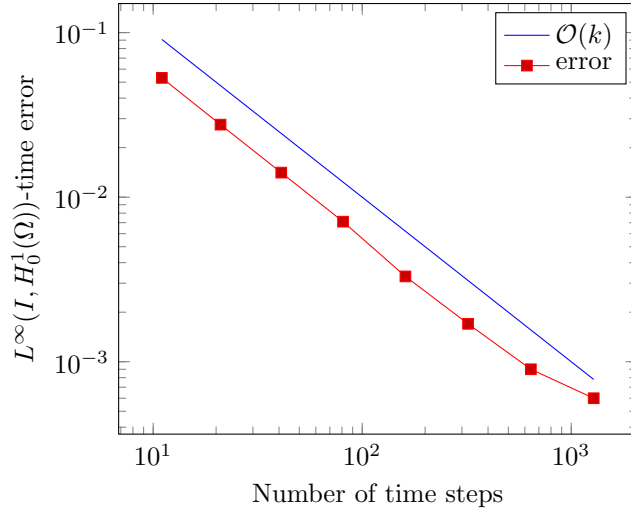


Figure 4.1: Temporal error for (4.182) with $N = 1\,050\,625$ spatial unknowns

For the error in time, we consider a sequence of discretizations having decreasing time steps with a fixed spatial triangulations with $N = 1\,050\,625$ spatial unknowns. The evolution of the error compared with the expected order of convergence $\mathcal{O}(k)$ is summarized in Figure 4.1. In Figure 4.2, we consider the error

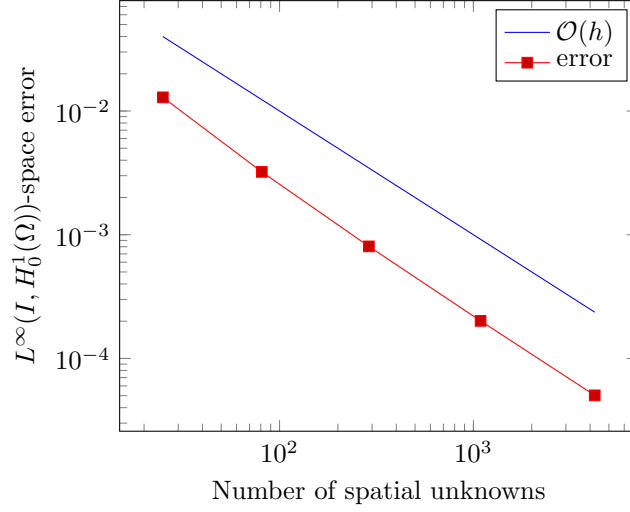


Figure 4.2: Spatial error for (4.182) for $M = 10\,001$ time steps

under refinements of the spatial triangulations for fixed times steps $M = 10\,001$. The error is compared with the expected order of convergence $\mathcal{O}(h)$. In both cases, the numerical results confirms the theoretical findings.

For the semi-linear state equation, we consider a similar problem with a quadratic non-linear term. Namely

$$\begin{aligned}
 (4.183) \quad & \partial_t u(t, x, y) - \Delta u(t, x, y) + u^2(t, x, y) = q(t)g(x, y) + f(t, x, y) && \text{in } (0, T) \times \Omega, \\
 & u(0, \cdot) = u_0 && \text{in } \Omega, \\
 & u(t, x, y) = 0 && \text{on } (0, T) \times \partial\Omega,
 \end{aligned}$$

with $I = (0, 1)$ and $\Omega = (0, \pi)^2$. The following data are chosen

$$\begin{aligned}
 q(t) &= (3 - 2t)e^{t-t^2}, \\
 g(x, y) &= \sin(x) \sin(y), \\
 u_0 &= \sin(x) \sin(y), \\
 f(t, x, y) &= e^{2t-2t^2} (\sin(x) \sin(y))^2,
 \end{aligned}$$

with the known solution given by

$$u(t, x, y) = e^{t-t^2} \sin(x) \sin(y).$$

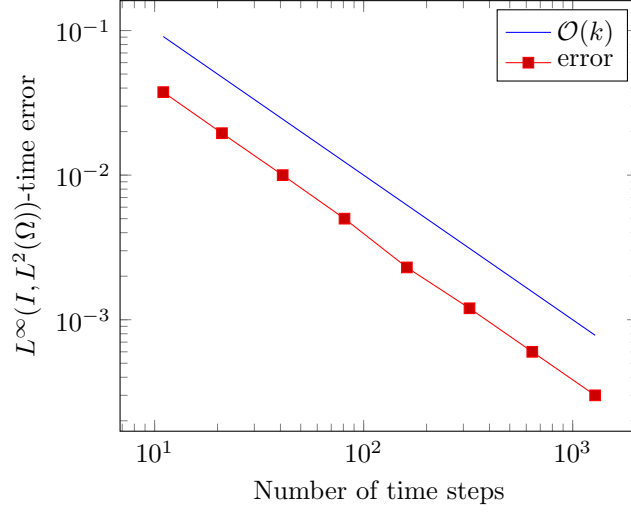


Figure 4.3: Temporal error for (4.183) with $N = 1\,050\,625$ spatial unknowns

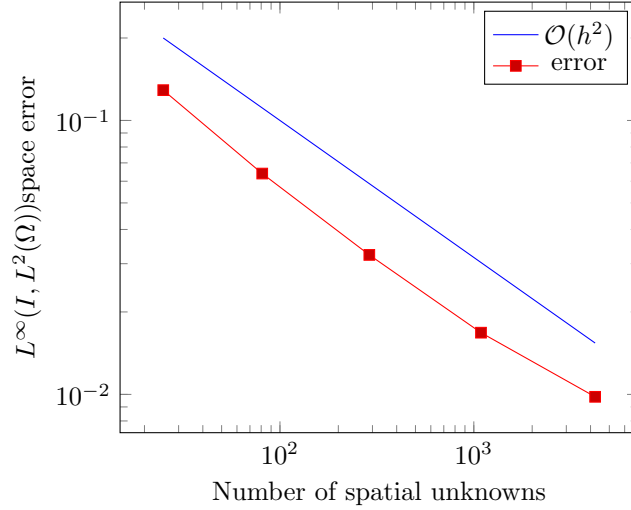


Figure 4.4: Spatial error for (4.183) for $M = 10\,001$ time steps

Also for the semi-linear case (4.183), our theoretical findings are confirmed by the numerical simulation. Figure 4.3 exhibits the evolution of the time error with a fixed spatial discretization of $N = 1\,050\,625$ spatial unknowns. Figure 4.4 shows the expected order of convergence $\mathcal{O}(h^2)$ compared with the error under refinement of the spatial discretization for fixed times steps $M = 10\,001$.

Finally, Figure 4.5 and 4.6 show the behavior of the temporal and spatial error, respectively, in the $L^\infty(I, H_0^1(\Omega))$ -norm for the semi-linear example (4.183). This agrees with our conjecture of Section 4.4.3 of an expected order of convergence $\mathcal{O}(k + h)$.

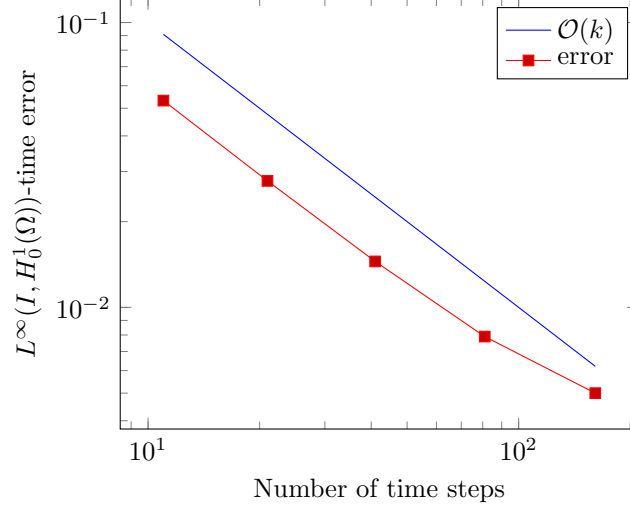


Figure 4.5: Temporal error for (4.183) with $N = 1\,050\,625$ spatial unknowns

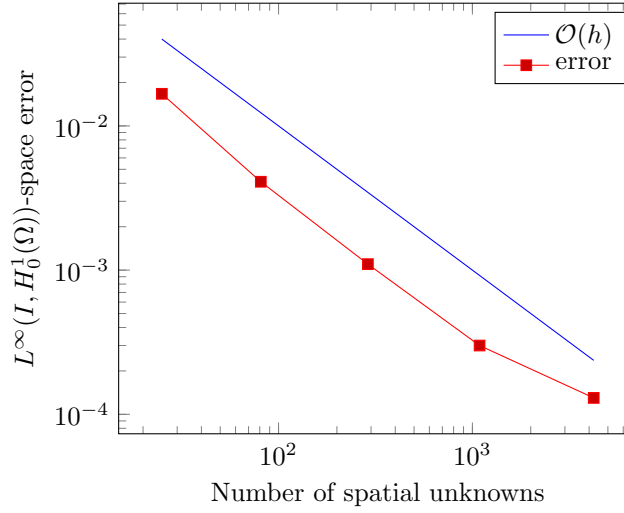


Figure 4.6: Spatial error for (4.183) for $M = 10\,001$ time steps

5. Optimization problems with state constraints

This chapter is devoted to the derivation of convergence rates for the error arising in the discretization of the optimal control problems (2.20) and (2.27). For this task, we exploit the error estimates for the state equation, derived in Chapter 4, together with the optimality conditions which we will hereafter derive in Section 5.1.

The approach used varies depending on the nature of the problem at hand, convex and non-convex, and it has to deal with the presence of the state constraint. Indeed, as anticipated in the Introduction and in Section 3.3, the Lagrange multiplier corresponding to the state constraint is a Borel measure whose presence as data in the equation defining the adjoint variable afflicts the regularity of the adjoint itself. Then it is clear that the derivation of convergence rate has to avoid the use of adjoint information.

In Problem 2.20 this is performed using a well-known procedure inherited from the elliptic setting, see, e.g., [43, Section 3.3], where the variational inequality is used in combination with the complementary slackness condition. Since the problem is convex, this strategy is legitimate because the first order conditions are necessary and sufficient for optimality. Following the approach of [61], in Section 5.2.1, we obtain a clear separation for the error in the time and space discretization.

For Problem 2.27 the situation is more complex due to its non-convex nature. Second order sufficient conditions (SSCs) need to be postulated in a suitable cone of critical directions. Such SSCs permit the derivation of a quadratic-growth condition which is the base to obtain the convergence of the optimal control problem. The difficulties introduced by the state constraint in this case are circumvented using the so-called *two-way feasibility*, dated back to [31] and reformulated in a modern fashion in [64] for an elliptic problem. Another issue in the non-convex setting is the presence of local solutions which requires the introduction of auxiliary localized problems, see [14]. The combination of this two arguments has been recently used in [66] for a semi-linear elliptic problem and we intend to extend it in Section 5.3 to the semi-linear parabolic setting.

We mention that for the non-convex case the convergence analysis will be done in one step, without analyzing separately the error arising from the time and space discretization. Indeed, this would require the transfer of SSCs to the time-discrete level which falls outside the scopes of this work.

The main result for Problem 2.20 reads

$$\|\bar{q} - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)}^2 \leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h \right),$$

and it will be shown in Theorem 5.2.1. For Problem 2.27 we obtain in Theorem 5.3.1

$$\|\bar{q} - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)}^2 \leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right).$$

We start with a discussion of optimality conditions in Section 5.1. Then, we analyze the convergence of the convex problem in Section 5.2.1. The convergence for the non-convex case is performed in Section 5.3. The material of Sections 5.2.1 and 5.3.1 has been obtained by the author of this thesis in [56, Section 5] and [55, Section 5], respectively.

5.1 Optimality conditions

We specify the Banach spaces introduced in the definition of the abstract optimization Problem 2.5 as

$$\mathcal{Q} = L^2(I, \mathbb{R}^m), \quad \mathcal{Z} = C(\bar{I}),$$

with the set of admissible controls given by

$$\mathcal{Q}_{\text{ad}} = \{q \in L^2(I, \mathbb{R}^m) \mid q_{\min} \leq q \leq q_{\max}\},$$

and the closed convex cone defined by

$$\mathcal{K} := \{v \in C(\bar{I}) \mid v - b \leq 0, \text{ in } \bar{I}\}.$$

We have seen in Sections 3.1 and 3.2 that the solution operators associated with the linear state equation (2.22) and the semi-linear one (2.28) are continuous. Thus, the well-posedness of the reduced cost functional

$$j(q) := J(q, S(q))$$

is guaranteed for both cases.

Further, we recall that the additional regularity of the solution of (2.22), given in Proposition 3.1.1, namely

$$U = L^2(I, H_0^1(\Omega) \cap H^2(\Omega)) \cap L^\infty(I, H_0^1(\Omega)) \cap H^1(I, L^2(\Omega)),$$

ensures that it holds the embedding $U \hookrightarrow C(\bar{I}, H_0^1(\Omega))$. Similarly, for the semi-linear state equation (2.28), without invoking the additional regularity, we have $W(0, T) \hookrightarrow C(\bar{I}, L^2(\Omega))$.

This is what is needed to treat the state constraints, indeed in both cases the state constraint is well-defined having range in the space $C(\bar{I})$, that is, the space involved in the definition of \mathcal{K} . Formally, using same notation for the state constraint, for Problem 2.20 we have

$$F: U \rightarrow C(\bar{I}), \quad F(u) = (|\nabla u|^2, \omega),$$

and for Problem 2.27

$$F: W(0, T) \rightarrow C(\bar{I}), \quad F(u) = (u, \omega).$$

Further, we denote with

$$(5.1) \quad G = (F \circ S): \mathcal{Q}_{\text{ad}} \rightarrow C(\bar{I})$$

the concatenation of the control-to-state map and the state constraint. It will be clear from the context whether G is referring to the convex or non-convex problem.

We now treat separately the convex and non-convex case.

5.1.1 First order optimality conditions

The convex case

In a first step, we postulate a Slater's regularity condition for Problem 2.20.

Assumption 5.1.1. *There exists $q_\gamma \in Q_{ad}$ such that*

$$(5.2) \quad G(q_\gamma) - b \leq -\gamma < 0$$

for some $\gamma \in \mathbb{R}^+$.

Proposition 5.1.1. *Under Assumption 5.1.1, Problem 2.20 admits a unique solution $\bar{q} \in L^\infty(I, \mathbb{R}^m)$ with corresponding state $\bar{u} = u(\bar{q}) \in U$.*

Proof. See the discussion in Theorem 2.2.4. The additional regularity is a consequence of the box constraints on the control variable. \square

We observe that the regularity condition formulated above reads

$$G(q_\gamma) \in \text{int } \mathcal{K},$$

coinciding with the one formulated in Theorem 2.2.5 for the abstract Problem 2.5. As we have seen, this entails the existence of a Lagrange multiplier satisfying, together with the optimal control \bar{q} , the variational inequality (2.9) and the complementary slackness condition (2.10).

In a next step, we write explicitly the optimality system in KKT-form introducing an adjoint state and using the Lagrangian approach, see, e.g., [43, Section 1.6.4].

Theorem 5.1.2. *Under Assumption 5.1.1 the pair $(\bar{q}, \bar{u}) \in Q_{ad} \times U$ is optimal for (2.20) if and only if it is feasible and there exists a Lagrange multiplier $\bar{\mu} \in C(\bar{I})^*$ and an adjoint state $\bar{z} \in L^2(I \times \Omega) \cap L^\infty(I, H^{-1}(\Omega))$ satisfying the following system of optimality conditions:*

$$\begin{aligned} (5.3a) \quad & (\partial_t \bar{u}, \varphi)_I + (\nabla \bar{u}, \nabla \varphi)_I = (\bar{q}g, \varphi)_I + (u_0, \varphi(0)) & \forall \varphi \in U, \\ (5.3b) \quad & (\partial_t \varphi, \bar{z})_I + (\nabla \varphi, \nabla \bar{z})_I = (\bar{u} - u_d, \varphi)_I + \langle 2(\nabla \bar{u} \nabla \varphi, \omega), \bar{\mu} \rangle & \forall \varphi \in U, \\ (5.3c) \quad & \alpha(\bar{q}, q - \bar{q})_{L^2(I)} + (\bar{z}, (q - \bar{q})g)_I \geq 0 & \forall q \in Q_{ad}, \\ (5.3d) \quad & \langle b - F(\bar{u}), \bar{\mu} \rangle = 0, \quad \bar{\mu} \geq 0, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $C(\bar{I})^*$ and $C(\bar{I})$.

Proof. We define the Lagrangian functional for Problem 2.20

$$\mathcal{L}(q, u(q), z, \mu): Q_{ad} \times U \times L^2(I \times \Omega) \times C(\bar{I})^* \rightarrow \mathbb{R}$$

as

$$(5.4) \quad \mathcal{L}(q, u(q), z, \mu) = J(q, u(q)) + (qg, z)_I - (\partial_t u, z)_I - (\nabla u, \nabla z) + \langle F(u(q)), \mu \rangle$$

using the formalism of having both the control q and the state u appearing in it. Then, the adjoint equation is given by the directional derivative of (5.4) with respect to the state variable, set equal to zero and evaluated at the optimal pair, namely

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{q}, \bar{u}, \bar{z}, \bar{\mu}) = 0.$$

With easy computations, we obtain (5.3b), whose solvability is guaranteed by the discussion in Section 3.3, noting that the right-hand side of (3.23) in our setting reads

$$\int_0^T v \, d\mu = \langle v, \mu \rangle, \quad v \in \mathcal{K}.$$

Similarly, the variational inequality is obtained taking the directional derivative with the respect to the control variable

$$\frac{\partial \mathcal{L}}{\partial q}(\bar{q}, \bar{u}, \bar{z}, \bar{\mu}) \geq 0$$

leading to (5.3c).

At last, equation (5.3d) is the usual complementary slackness condition and (5.3a) expresses the feasibility of (\bar{q}, \bar{u}) . \square

Remark 5.1.3. *For the derivation of the KKT-system, we have used the formal Lagrangian method which has the advantage to make the exposition simpler, avoiding the specification of the underlying functional spaces and the introduction of cumbersome notations. For an application of the exact Lagrange method in a context similar to the one at hand, we refer to [85, Chapter 6], see also [61, Theorem 2.4].*

After deriving the KKT-system for the continuous problem, we show that the Slater's condition continues to hold also for the semi and fully discrete problems and we derive the corresponding KKT-systems. In the following, we denote with

$$(5.5) \quad G_k: Q_{\text{ad}} \rightarrow U_k,$$

the concatenation of the solution operator of (2.23) and the state constraint (2.20d), with U_k defined in (2.14).

Lemma 5.1.4. *For the control $q_\gamma \in Q_{\text{ad}}$ satisfying Assumption 5.1.1 there holds*

$$(5.6) \quad G_k(\tilde{q}) - b \leq -\frac{\gamma}{2} < 0$$

for $\gamma \in \mathbb{R}^+$ and for k sufficiently small.

Proof. Using Assumption 5.1.1 and by virtue of Theorem 4.3.6, we have

$$\begin{aligned} F(u_k(\tilde{q})) &= F(u(\tilde{q})) + F(u_k(\tilde{q}) - u(\tilde{q})) \\ &< b - \gamma + \|\omega\|_{L^\infty(\Omega)} \|u(\tilde{q}) - u_k(\tilde{q})\|_{L^\infty(I, H_0^1(\Omega))} \\ &< b - \gamma + Ck \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \end{aligned}$$

and the claim follows once k is sufficiently small. \square

We recall that the definition of the bilinear form $B(\cdot, \cdot)$ has been given in (2.15).

5.1. Optimality conditions

Theorem 5.1.5. *Under Assumption 5.1.1 the pair $(\bar{q}_k, \bar{u}_k) \in Q_{ad} \times U_k$ is optimal for (2.25) if and only if it is feasible and there exists a Lagrange multiplier $\bar{\mu}_k \in C(\bar{I})^*$ and an adjoint state $\bar{z}_k \in U_k$ satisfying the following system of optimality conditions*

$$\begin{aligned} (5.7a) \quad & B(\bar{u}_k, \varphi) = (\bar{q}_k g, \varphi)_I + (u_0, \varphi_0^+) & \forall \varphi \in U_k, \\ (5.7b) \quad & B(\varphi, \bar{z}_k) = (\bar{u}_k - u_d, \varphi)_I + \langle 2(\nabla \bar{u}_k \nabla \varphi, \omega), \bar{\mu}_k \rangle & \forall \varphi \in U_k, \\ (5.7c) \quad & \alpha(\bar{q}_k, q - \bar{q}_k)_{L^2(I)} + (\bar{z}_k, (q - \bar{q}_k)g)_I \geq 0 & \forall q \in Q_{ad}, \\ (5.7d) \quad & \langle b - F(\bar{u}_k), \bar{\mu}_k \rangle = 0, \end{aligned}$$

where the Lagrange multiplier $\bar{\mu}_k$ is given by

$$(5.8) \quad \langle \bar{\mu}_k, v \rangle = \sum_{n=1}^N \frac{\mu_{k,n}}{k_n} \int_{I_n} v(t) dt, \quad \forall v \in C(\bar{I}) \cup U_k(\mathbb{R})$$

with $\mu_{k,n} \in \mathbb{R}^+$ for any $n = 1, \dots, N$.

Proof. We observe that the state constraint in the semi-discrete setting reads

$$F(u_k)|_{I_n} \leq b, \quad \text{for } n = 1, \dots, N$$

due to u_k being piecewise constant in time.

This in particular means that the convex cone of non-positive function has to be modified to include the presence of finitely many state constraints. Therefore we define

$$\mathcal{K} = \{v \in \mathbb{R}^N \mid v_n \leq 0, n = 1, \dots, N\}$$

and we rewrite Problem 2.25 as

$$\min j_k(q) \text{ s.t. } q \in Q_{ad}, G_k(q) \in \mathcal{K}.$$

Thanks to Lemma 5.1.4 and with same argument as in the continuous case, we have the existence of finitely many Lagrange multipliers $\mu_{k,n} \in \mathbb{R}_+$ for all $n = 1, \dots, N$, associated to the subintervals I_n . This leads to $\bar{\mu}_k \in C(\bar{I})^*$ by construction in (5.8). Then, we obtain the optimality system proceeding as in Theorem 5.1.2. \square

Repeating the steps of Lemma 5.1.4 with Theorem 4.3.13 in place of Theorem 4.3.6, one obtains that Assumption 5.1.1 continues to hold for the discrete problem. Then, arguing as in Theorem 5.1.5 we infer the optimality system for Problem 2.26.

Theorem 5.1.6. *Under Assumption 5.1.1 the pair $(\bar{q}_{kh}, \bar{u}_{kh}) \in Q_{ad} \times U_{kh}$ is optimal for (2.26) if and only if it is feasible and there exists a Lagrange multiplier $\bar{\mu}_{kh} \in C(\bar{I})^*$ and an adjoint state $\bar{z}_{kh} \in U_{kh}$ satisfying the following system of optimality conditions*

$$\begin{aligned} (5.9a) \quad & B(\bar{u}_{kh}, \varphi) = (\bar{q}_{kh} g, \varphi)_I + (u_0, \varphi_0^+) & \forall \varphi \in U_{kh}, \\ (5.9b) \quad & B(\varphi, \bar{z}_{kh}) = (\bar{u}_{kh} - u_d, \varphi)_I + \langle 2(\nabla \bar{u}_{kh} \nabla \varphi, \omega), \bar{\mu}_{kh} \rangle & \forall \varphi \in U_{kh}, \\ (5.9c) \quad & \alpha(\bar{q}_{kh}, q - \bar{q}_{kh})_{L^2(I)} + (\bar{z}_{kh}, (q - \bar{q}_{kh})g)_I \geq 0 & \forall q \in Q_{ad}, \\ (5.9d) \quad & \langle b - F(\bar{u}_{kh}), \bar{\mu}_{kh} \rangle = 0, \end{aligned}$$

where the Lagrange multiplier $\bar{\mu}_{kh}$ is given by

$$(5.10) \quad \langle \bar{\mu}_{kh}, v \rangle = \sum_{n=1}^N \frac{\mu_{kh,n}}{k_n} \int_{I_n} v(t) dt, \quad \forall v \in C(\bar{I}) \cup U_k(\mathbb{R})$$

with $\mu_{kh,n} \in \mathbb{R}^+$ for any $n = 1, \dots, N$.

We remark again that being Problem 2.20 convex, the first-order conditions are necessary and sufficient for optimality.

The non-convex case

Guided by the discussion in the previous paragraph, we formulate the first-order necessary conditions for the non-convex case. In a first step, we infer the existence of a local solution for this problem.

Proposition 5.1.7. *Under Assumption 2.3.1, Problem 2.27 admits at least a local solution $\bar{q} \in Q_{ad}$ with corresponding state $\bar{u} \in W(0, T)$.*

Proof. The proof follows the lines of Theorem 2.2.4. In particular, let $\{q_n\} \subset Q_{ad}$ be a minimizing sequence and let $u(q_n)$ be the corresponding sequence of state variables. As in Theorem 2.2.4, we can extract from $\{q_n\}$ a weakly convergent subsequence with limit \bar{q} . Further, defining a linear parabolic problem having as data q_n and $-d(t, x, u(q_n))$, we infer that $u(q_n) \rightharpoonup \bar{u}$, invoking the continuity of the control-to-state map. So far, the candidate for optimality is \bar{q} and consequently $u(\bar{q})$. Thus, we must show that $u(\bar{q}) = \bar{u}$, that is, the sequence $u(q_n)$ converges to the right state variable. For this we need the strong convergence in a suitable space of the Nemytskii operator, namely

$$d(t, x, u(q_n)) \rightarrow d(t, x, \bar{u}), \quad \text{in } C(\bar{I} \times \bar{\Omega})$$

which is guaranteed by Proposition 3.2.1.

Then, the (local) optimality of \bar{q} follows as in the convex case exploiting the weak-lower semi-continuity of j . \square

The Slater's regularity condition from Assumption 5.1.1 is now linearized in a standard manner to deal with the presence of the semi-linear term in the state equation.

Assumption 5.1.2. *Given a local solution \bar{q} of Problem 2.27, there exists $q_\gamma \in Q_{ad}$ such that*

$$(5.11) \quad G(\bar{q}) + G'(\bar{q})(q_\gamma - \bar{q}) - b < -\gamma < 0$$

for $\gamma \in \mathbb{R}^+$.

Remark 5.1.8. *Evaluating (5.11) at $t = 0$, we have a further condition to assume on the initial data u_0 , namely*

$$F(u_0) - b < -\gamma < 0, \quad \text{on } \bar{\Omega}.$$

With a regularity condition at hand, the derivation of the KKT-optimality system follows along the lines of Theorem 5.1.2.

Theorem 5.1.9. *Under Assumption 2.3.1, let $\bar{q} \in Q_{ad}$ be a local solution for Problem 2.27 satisfying 5.1.2 the pair, and let $\bar{u} \in W(0, T)$ be the associated state. Then, there exists a Lagrange multiplier $\bar{\mu} \in C(\bar{I})^*$ and an adjoint state $\bar{z} \in L^2(I, H_0^1(\Omega)) \cap L^\infty(I, L^2(\Omega))$ such that*

$$\begin{aligned} (5.12a) \quad & b(\bar{u}, \varphi) + (d(\cdot, \cdot, \bar{u}), \varphi)_I = (\bar{q}g, \varphi)_I + (u_0, \varphi(0)) \quad \forall \varphi \in W(0, T), \\ (5.12b) \quad & b(\varphi, \bar{z}) + (\varphi, \partial_u d(\cdot, \cdot, \bar{u})\bar{z}) = (\bar{u} - u_d, \varphi)_I + \langle F(\varphi), \bar{\mu} \rangle \quad \forall \varphi \in W(0, T), \\ (5.12c) \quad & \alpha(\bar{q}, q - \bar{q})_{L^2(I)} + (\bar{z}, (q - \bar{q})g)_I \geq 0 \quad \forall q \in Q_{ad}, \\ (5.12d) \quad & \langle b - F(\bar{u}), \bar{\mu} \rangle = 0, \quad \bar{\mu} \geq 0, \quad F(\bar{u}) \leq 0 \end{aligned}$$

where we used the linearity of $F(\cdot)$, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $C(\bar{I})^*$ and $C(\bar{I})$, and $b(\cdot, \cdot)$ is the bilinear form defined in (2.21).

Proof. See [20, Theorem 4] and the reference therein. \square

As the convergence of Problem 2.27 will be performed in one step, from the continuous to the discrete level, we do not need explicitly the KKT-system for the time-discrete problem. Therefore we formulate directly the KKT optimality conditions for the the discrete problem. These conditions will be justified after the introduction of an auxiliary problem in Section 5.3, which will guarantee that the Slater point for Assumption 5.1.2 is also a Slater point for the discrete problem.

Theorem 5.1.10. *Let $\bar{u}_{kh} \in Q_{kh,feas}$ be a local solution of Problem 2.32 with $\bar{u}_{kh} \in U_{kh}$ the associated state. Then, under Assumption 5.1.2, for k, h sufficiently small there exists a Lagrange multiplier $\bar{\mu}_{kh} \in U_{kh}(\mathbb{R})^* \cap C(\bar{I})^*$ and an adjoint state $\bar{z}_{kh} \in U_{kh}$ such that*

$$\begin{aligned} B(\bar{u}_{kh}, \varphi) + (d(\cdot, \cdot, \bar{u}_{kh}), \varphi)_I &= (\bar{q}_{kh}g, \varphi)_I + (u_0, \varphi_{kh,1}) \quad \forall \varphi \in U_{kh}, \\ B(\varphi, \bar{z}_{kh}) + (\varphi, \partial_u d(\cdot, \cdot, \bar{u}_{kh})\bar{z}_{kh}) &= (\bar{u} - u_d, \varphi)_I + \langle F_{kh}(\varphi), \bar{\mu}_{kh} \rangle \quad \forall \varphi \in U_{kh}, \\ \alpha(\bar{q}_{kh}, q - \bar{q}_{kh})_{L^2(I)} + (\bar{z}_{kh}, (q - \bar{q}_{kh})g)_I &\geq 0 \quad \forall q \in Q_{kh,feas}, \\ \langle F_{kh}(\bar{u}_{kh}), \bar{\mu}_{kh} \rangle &= 0, \quad \bar{\mu} \geq 0, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $U_{kh}(\mathbb{R})^*$ and $U_{kh}(\mathbb{R})$. Further, the Lagrange multiplier can be represented as an element of $C(\bar{I})^*$ by

$$\langle v, \bar{\mu}_{kh} \rangle = \sum_{n=1}^N \frac{\mu_{kh,n}}{k_n} \int_{I_n} v(t) dt, \quad \forall v \in C(\bar{I}) \cup U_{kh}(\mathbb{R}).$$

From now on, with no loss of generality, we will set the bound in the integral state constraint (2.27d) $b = 0$. This is for sake of readability, as in the following sections, we will introduce further notation, and for an easy comparison with the author's paper [55] where this problem is studied.

5.1.2 Second order optimality conditions

This section is devoted to the derivation of the SSCs for Problem 2.27. We adopt a recent approach exposed in [16] which makes use of an application of Egorov's theorem in the context of optimal control. The cone of critical directions which will be used, is inspired by the one firstly introduced in [8] and extended in [20]

for a setting similar to the one at hand.

In a first step, we state a result, see [16, Lemma 3.5], which we will comment later.

Lemma 5.1.11. *Let (X, Σ, μ) be a measure space of finite measure and assume the existence of $\{f_k\}_{k=1}^\infty \subset L^\infty(X)$, bounded in $L^\infty(X)$ with $f_k \geq 0$ a.e. in X and $f_k \rightarrow f$ in $L^1(X)$. Further, let $\{p_k\}_{k=1}^\infty \subset L^2(X)$ be a sequence converging weakly in $L^2(X)$ to some p . Then it holds*

$$(5.13) \quad \int_X f p^2 d\mu \leq \liminf_{k \rightarrow \infty} \int_X f_k p_k^2 d\mu.$$

Proposition 5.1.12. *Let $\{q_k\}_{k=1}^\infty \subset Q_{ad}$ and $\{p_k\} \subset L^2(I, \mathbb{R}^m)$ be such that*

$$q_k \rightarrow q \text{ in } L^2(I, \mathbb{R}^m) \quad p_k \rightharpoonup p \text{ in } L^2(I, \mathbb{R}^m).$$

Then the following relations holds true

$$(5.14) \quad j''(q)p^2 \leq \liminf_{k \rightarrow \infty} j''(q_k)p_k^2$$

$$(5.15) \quad \text{if } p = 0, \quad \Lambda \liminf_{k \rightarrow \infty} \|p_k\|_{L^2(I, \mathbb{R}^m)}^2 \leq \liminf_{k \rightarrow \infty} j''(q_k)p_k^2$$

for some $\Lambda > 0$.

Proof. Recalling the expression for the second derivative of j from Corollary 3.4.5,

$$(5.16) \quad j''(q)p_1 p_2 = \int_I \int_\Omega (v_{p_1} v_{p_2} - z_0(q) \partial_u^2 d(x, t, u(q)) v_{p_1} v_{p_2}) dx dt + \int_I p_1^T \alpha p_2 dt,$$

for relation (5.14) we have to analyze the convergence of $z_0(q_k)$, defined by (3.33) and the one of v_{p_k} defined by (3.29). To do so, we first observe that Proposition 3.2.2 ensures that $S(q_k) \rightarrow S(q)$ in $L^2(I, H_0^1(\Omega)) \cap C(\bar{I} \times \Omega)$. This and the boundedness of $\partial_u d(\cdot, \cdot, u)$, implies that the same holds true for the solution of (3.33), that is, $z_0(q_k) \rightarrow z_0(q)$ in $L^2(I, H_0^1(\Omega)) \cap C(\bar{I} \times \Omega)$. Similarly, we have the convergence $v_{p_k} = S'(q_k)v_k \rightarrow S'(q)v = v_p$ in $L^2(I, H_0^1(\Omega)) \cap C(\bar{I} \times \Omega)$. These show (5.14).

Last relation easily follows observing that, being $p = 0$, it is enough to set $\Lambda = \alpha$. \square

Remark 5.1.13. *The result above is simplified by the setting of our problem. In particular, the use of Lemma 5.1.11 is hidden by the structure of the objective functional where the control appears quadratically. To highlight the use of Lemma 5.1.11, let assume that the objective functional is defined in an abstract way by a function $\varphi(t, x, q, u)$ defined as in [85, Assumption 5.6]. Then in the expression for $j''(q)p^2$ we would have to deal with the term $p^T \partial_q^2 \varphi p$ which in our setting is reduced to $p^T \alpha p$.*

Then relation (5.14) follows applying Lemma 5.1.11 with

$$f_k = \partial_q^2 \varphi(t, x, u(q_k), q_k), \text{ and } f = \partial_q^2 \varphi(t, x, u(q), q),$$

and $X = I \times \Omega$, with μ be the corresponding Lebesgue measure.

Further, to show relation (5.15) in a more general setting, one has to impose a so-called Legendre-Clebsch condition

$$\exists \Lambda > 0 \text{ s.t. } \partial_q^2 \varphi(t, x, q, u) \geq \Lambda \text{ for a.a. } (t, x) \in I \times \Omega \text{ and } \forall u, q \in \mathbb{R},$$

5.1. Optimality conditions

see [16, Equation (5.3) and Proposition 5.3].

For the discussion of the second order sufficient optimality conditions, we introduce the Hamiltonian function associated with Problem 2.27

$$H(t, x, q, u, z): I \times \Omega \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$$

given by

$$(5.17) \quad H(t, x, q, u, z) = H(q, u, z) = \frac{1}{2}(u - u_d)^2 + \frac{\alpha}{2}q^2 + z \left(\sum_{i=1}^m q_i g_i - d(u) \right),$$

and the reduced Lagrangian function

$$(5.18) \quad \mathcal{L}(q, \mu) = j(q) + \langle G(q), \mu \rangle.$$

When no confusion arises, we denote the Hamiltonian and Lagrangian evaluated at $(\bar{q}, \bar{u}, \bar{z}, \bar{\mu})$ for each (t, x) with $\bar{H}, \bar{\mathcal{L}}$. Being $\frac{\partial H}{\partial q}$ a \mathbb{R}^m -vector, we denote with $\frac{\partial H_i}{\partial q}$ its i -th component. Same holds for the (i, j) -entry of the $\mathbb{R} \times \mathbb{R}$ matrix $\frac{\partial^2 H_{i,j}}{\partial q^2}$. Further, for sake of readability, in the following, for a $p \in L^2(I, \mathbb{R}^m)$, we will use

$$\frac{\partial^2 \mathcal{L}}{\partial q^2} p^2 \quad \text{instead of} \quad p^T \frac{\partial^2 \mathcal{L}}{\partial q^2} p.$$

We are now ready to formulate the cone of critical direction associated with a feasible control $\bar{q} \in Q_{ad}$. For this we introduce the conditions

$$(5.19) \quad p_i(t) = \begin{cases} \geq 0 & \text{if } \bar{q}_i = q_{min}, \\ \leq 0 & \text{if } \bar{q}_i = q_{max}, \\ = 0 & \text{if } \int_{\Omega} \frac{\partial \bar{H}_i}{\partial q} dx \neq 0, \end{cases} \quad \text{for all } i = 1, \dots, m$$

$$(5.20) \quad \frac{\partial F}{\partial u}(\bar{u}) v_p \leq 0 \text{ if } F(\bar{u}) = 0,$$

$$(5.21) \quad \int_{\mathcal{K}} \frac{\partial F}{\partial u}(\bar{u}) v_p d\bar{\mu} = 0,$$

where v_p is defined by (3.29). Then, the cone of critical directions is defined by

$$(5.22) \quad C_{\bar{q}} = \{p \in L^2(I, \mathbb{R}^m) \mid p \text{ satisfies (5.19), (5.20), (5.21)}\}.$$

We now state the SSCs and derive a quadratic growth condition based on it.

Assumption 5.1.3. *Let $\bar{q} \in Q_{ad}$ be a feasible control fulfilling together with the associated state \bar{u} , the adjoint state \bar{z} , and Lagrange multiplier $\bar{\mu}$ the first order necessary conditions (5.12). Then we assume*

$$(5.23) \quad \frac{\partial^2 \bar{\mathcal{L}}}{\partial q^2} p^2 > 0, \quad \forall p \in C_{\bar{q}} \setminus \{0\}.$$

Theorem 5.1.14. *Under Assumption 5.1.3, let $\bar{q} \in Q_{ad}$ be a feasible control satisfying the first order necessary optimality conditions (5.12). Then there exists constants $\delta, \eta > 0$ such that*

$$(5.24) \quad j(q) \geq j(\bar{q}) + \delta \|q - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2$$

for any feasible control of Problem 2.27 with $\|q - \bar{q}\|_{L^2(I, \mathbb{R}^m)} \leq \eta$.

Proof. The proof is based on a contradiction argument introduced in [8, Theorem 4.1] and extended in [20, Theorem 5] for the case of time-dependent controls. Our proof differs mainly in the construction of the final contradiction where we exploit (5.14) and (5.15), compare with [16, Theorem 2.3].

In a first step, we extend (5.14) and (5.15), formulated for the second derivative of the objective functional j , to the second derivative of the Lagrangian \mathcal{L} . From the definition of the Lagrangian (5.18), we observe that, for directions $p_1, p_2 \in L^2(I, \mathbb{R})$,

$$j''(q)p_1p_2 \quad \text{and} \quad \frac{\partial^2 \mathcal{L}}{\partial q^2} p_1p_2$$

only differ for the presence in the former of z_0 solution of (3.33), while in the latter this is substituted by the full adjoint variable z solution of (5.12b). This is the case because the state constraint (2.27d) is of zero order and, therefore, its second derivative with respect the control variable is zero. Then, since in the following arguments the Lagrange multiplier $\bar{\mu}$ remains fixed, we infer that relations (5.14) and (5.15) continue to hold for the Lagrangian, namely

$$(5.25) \quad \frac{\partial^2 \mathcal{L}}{\partial q^2}(q)p^2 \leq \liminf_{k \rightarrow \infty} \frac{\partial^2 \mathcal{L}}{\partial q^2}(q_k)p_k^2$$

$$(5.26) \quad \text{if } p = 0, \quad \Lambda \liminf_{k \rightarrow \infty} \|p_k\|_{L^2(I, \mathbb{R})}^2 \leq \liminf_{k \rightarrow \infty} \frac{\partial^2 \mathcal{L}}{\partial q^2}(q_k)p_k^2,$$

where p, q, p_k , and q_k are defined as in Proposition 5.1.12. This follows using same the steps as in Proposition 5.1.12 with the difference given by the convergence of $z(q_k)$ toward $z(q)$ in $L^2(I, H_0^1(\Omega)) \cap L^\infty(I, L^2(\Omega))$, which is guaranteed by (3.25), see also [7, Theorem 6.4].

After this preamble, we now construct the contradiction. Assuming that \bar{q} does not satisfy (5.24), then there exists a sequence of controls $\{q_k\}_{k=1}^\infty \subset Q_{ad}$ feasible for (2.27) such that

$$\|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)} < \frac{1}{k}$$

and

$$(5.27) \quad j(q_k) < j(\bar{q}) + \frac{1}{2k} \|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2.$$

In a next step, we build a direction which will be used to obtain the final contradiction. We set

$$\rho_k = \|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)}, \quad \text{and} \quad p_k = \frac{(q_k - \bar{q})}{\rho_k}.$$

5.1. Optimality conditions

Noting that $\|p_k\|_{L^2(I, \mathbb{R}^m)} = 1$, we extract a subsequence, using same notation, such that

$$p_k \rightharpoonup p, \text{ in } L^2(I, \mathbb{R}^m).$$

With a classical procedure, we check in a first step that the derivative in the direction p of the Lagrangian evaluated at $(\bar{q}, \bar{\mu})$ is zero. Observing that the complementary slackness condition (5.12d) implies $\bar{\mathcal{L}} = j(\bar{q})$, we obtain from (5.27), and the feasibility of q_k in (5.12d), that

$$\begin{aligned} (5.28) \quad j(\bar{q}) + \frac{1}{2k} \|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2 &= \mathcal{L}(\bar{q}, \bar{\mu}) + \frac{1}{2k} \|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2 \\ &> j(q_k) \\ &\geq \mathcal{L}(q_k, \bar{\mu}) \\ &= \mathcal{L}(\bar{q}, \bar{\mu}) + \rho_k \frac{\partial \mathcal{L}}{\partial q}(\bar{q}_k, \bar{\mu}) p_k, \end{aligned}$$

using in the last step the mean value theorem, with \tilde{q}_k being a point between q_k and \bar{q} , compare with [8, Equation (4.8)]

Then, rearranging the terms above and using again $\bar{\mathcal{L}} = j(\bar{q})$, we have

$$\frac{\partial \mathcal{L}}{\partial q}(\tilde{q}_k, \bar{\mu}) p_k \leq \frac{1}{2k \rho_k} \|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2 = \frac{1}{2k} \|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)},$$

and therefore, in the limit k going to infinity, it follows

$$(5.29) \quad \frac{\partial \mathcal{L}}{\partial q}(\bar{q}, \bar{\mu}) p \leq 0.$$

On the other hand, the converse inequality is also true noting that the feasibility of q_k together with the variational inequality (5.12c) yield

$$\frac{\partial \mathcal{L}}{\partial q}(\bar{q}, \bar{\mu}) p_k = \frac{1}{\rho_k} \frac{\partial \mathcal{L}}{\partial q}(\bar{q}, \bar{\mu}) (q_k - \bar{q}) \geq 0.$$

Taking the limit, we have

$$(5.30) \quad \frac{\partial \mathcal{L}}{\partial q}(\bar{q}, \bar{\mu}) p \geq 0,$$

Hence, (5.29) and (5.30) imply

$$\frac{\partial \mathcal{L}}{\partial q}(\bar{q}, \bar{\mu}) p = 0.$$

The second step consists in showing that the direction at hand is critical, that is, $p \in C_{\bar{q}}$. This has been shown in [8, Theorem 4.1] and extended in [20, Theorem 5, Step 2] to the case of time-dependent controls. The proof employs fairly standard arguments and it is based on the formulation of the first-order optimality conditions as Pontryagin's principle. We omit the details referring to the references before-mentioned.

We now show that $p = 0$ which together with (5.26) will given the final contradiction. Proceeding in the first step as in (5.28), and using a Taylor expansion

afterward, we have, for $\theta_k \in (0, 1)$,

$$\begin{aligned} \frac{1}{2k} \|q_k - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2 &> \mathcal{L}(q_k, \bar{\mu}) - \mathcal{L}(\bar{q}, \bar{\mu}) \\ &= \rho_k \frac{\partial \mathcal{L}}{\partial q} p_k + \frac{\rho_k^2}{2} \frac{\partial^2 \mathcal{L}}{\partial q^2} (\bar{q} + \theta_k \rho_k p_k) p_k^2 \\ &\geq \frac{\rho_k^2}{2} \frac{\partial^2 \mathcal{L}}{\partial q^2} (\bar{q} + \theta_k \rho_k p_k) p_k^2, \end{aligned}$$

using in the last step the variational inequality expressed in the form given in (2.9).

Thus, we have shown that

$$\frac{\partial^2 \mathcal{L}}{\partial q^2} (\bar{q} + \theta_k \rho_k p_k) p_k^2 < \frac{1}{k}.$$

Using the expression above together with (5.25) and (5.23), we have

$$0 \leq \frac{\partial^2 \mathcal{L}}{\partial q^2} (q) p^2 \leq \liminf_{k \rightarrow \infty} \frac{\partial^2 \mathcal{L}}{\partial q^2} (\bar{q} + \theta_k \rho_k p_k) p_k^2 \leq \lim_{k \rightarrow 0} \frac{1}{k} = 0.$$

The expression above implies

$$\frac{\partial^2 \mathcal{L}}{\partial q^2} (q) p^2 = 0,$$

and, therefore, by means of (5.23), $p = 0$.

Ultimately, observing that by construction it holds $\|p_k\|_{L^2(I, \mathbb{R})}^2 = 1$, thanks to (5.26) we obtain the contradiction

$$0 < \Lambda = \Lambda \liminf_{k \rightarrow \infty} \|p_k\|_{L^2(I, \mathbb{R})}^2 \leq \liminf_{k \rightarrow \infty} \frac{\partial^2 \mathcal{L}}{\partial q^2} (\bar{q} + \theta_k \rho_k p_k) p_k^2 = 0.$$

□

Remark 5.1.15. *The proof above can be alternatively obtained using the method introduced in [8, Theorem 4.1] and extended in [20, Theorem 5] for the case of time-dependent controls. Comparing it with our proof, it differs in the construction of the final contradiction. Further, the fact that the control is linear in the state equation and quadratic in the cost functional lead to the formulation of the quadratic growth condition without two-norm discrepancy, compare with [8, Theorem 4.2].*

We conclude the section with a discussion of the SSCs and of the cone of critical directions. Comparing $C_{\bar{q}}$ defined by (5.22), with the one in [8], where it was firstly introduced, we note that, for $\xi, \nu > 0$, the presence of an additional assumption

$$\frac{\partial \bar{H}_{i,i}}{\partial q^2} \geq \xi, \quad \forall t \in I \setminus E_i^\nu, \forall i = 1, \dots, m,$$

where

$$E_i^\nu = \left\{ t \in I \text{ s.t. } \left| \int_\Omega \frac{\partial \bar{H}_i}{\partial q} dx \right| \geq \nu \right\}$$

denotes the set of sufficiently active control constraints. This condition is automatically satisfied in our setting because the control appears linearly in the state equation and quadratically in the cost functional. Therefore, from (5.17), we obtain that

$$\frac{\partial \bar{H}_{i,i}}{\partial q^2} = \alpha \mathbb{I} > 0$$

where \mathbb{I} is the identity operator.

Then, if we are able to show the following second order necessary condition

$$(5.31) \quad \frac{\partial^2 \bar{\mathcal{L}}}{\partial q^2} p^2 \geq 0, \quad \forall p \in C_{\bar{q}},$$

we would end up with *no-gap* second-order optimality conditions.

In order to obtain (5.31), one can proceed as in [15, Theorem 2.2], see also [11, Theorem 3.3] for the case of an elliptic equation and integral state constraints. Though being an interesting matter, this would fall beyond the scope of this work and therefore we omit it.

5.2 Convergence analysis for convex optimization problems

This section is concerned with the estimate of the error arising from the time-space discretization of the convex-problem (2.20). The material presented here has been exposed in [56, Section 5].

5.2.1 First-order integral constraints pointwise in time

Based on the a priori error estimates derived in Section 4.3, we analyze the convergence of \bar{q}_{kh} solution of (2.26) toward the continuous solution \bar{q} of (2.20). We consider the influence of the time and space discretization separately and we obtain the main result of this section in the following, whose proof will be given combining Theorems 5.2.3 and 5.2.5.

Theorem 5.2.1. *Let $\bar{q} \in Q_{ad}$ be the optimal control of the continuous problem (2.20) and $\bar{q}_{kh} \in Q_{ad}$ be the one of the discrete problem (2.26). Then there holds the following error estimate*

$$(5.32) \quad \|\bar{q} - \bar{q}_{kh}\|_{L^2(I)}^2 \leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h \right).$$

Proof. The proof will be concluded combining Theorems 5.2.3 and 5.2.5 where the convergence rates induced by the time and space discretization are analyzed separately. \square

In a first step, we show that the semi-discrete optimal pair and the corresponding Lagrange multiplier can be bounded independently from the discretization.

Lemma 5.2.2. *Let $(\bar{q}_k, \bar{u}_k) \in Q_{ad} \times U_k$ be the optimal pair for Problem 2.25 with associated Lagrange multiplier $\bar{\mu}_k \in C(\bar{I})^*$. Then, for k sufficiently small there holds*

$$(5.33) \quad \|\bar{q}_k\|_{L^2(I, \mathbb{R}^m)} + \|\bar{u}_k\|_I + \|\bar{\mu}_k\|_{C(\bar{I})^*} \leq C.$$

Proof. We refer to [61, Lemma 6.2] where the claim is shown in a setting with integral constraint point-wise in time on the state variable with relative estimate for the state equation in $L^\infty(I, L^2(\Omega))$ -norm. Clearly, the arguments used there continue to hold in our setting substituting the $L^\infty(I, L^2(\Omega))$ -norm estimate with the $L^\infty(I, H_0^1(\Omega))$ -norm estimate derived in Theorem 4.3.6. \square

We now give the first intermediate result regarding the convergence of the semi-discrete optimal control. The following corresponds to [56, Theorem 5.1].

Theorem 5.2.3. *Let $\bar{q} \in Q_{ad}$ be the optimal control of the continuous problem (2.20) and $\bar{q}_k \in Q_{ad}$ be the one of the semi-discrete problem (2.25). Then there holds the following error estimate*

$$(5.34) \quad \alpha \|\bar{q} - \bar{q}_k\|_{L^2(I), \mathbb{R}^m}^2 \leq C k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}}.$$

Proof. We consider the continuous and semi-discrete variational inequality (5.3c) and (5.7c) with $q = \bar{q}_k$ and $q = \bar{q}$, respectively, obtaining

$$\begin{aligned} 0 &\leq \alpha(\bar{q}, \bar{q}_k - \bar{q})_{L^2(I)} + ((\bar{q}_k - \bar{q})g, \bar{z})_I, \\ 0 &\leq \alpha(\bar{q}_k, \bar{q} - \bar{q}_k)_{L^2(I)} + ((\bar{q} - \bar{q}_k)g, \bar{z}_k)_I. \end{aligned}$$

By adding these inequalities, we have

$$\begin{aligned} \alpha \|\bar{q} - \bar{q}_k\|_{L^2(I), \mathbb{R}^m}^2 &= -\alpha(\bar{q} - \bar{q}_k, \bar{q}_k - \bar{q}) \\ &\leq ((\bar{q}_k - \bar{q})g, \bar{z} - \bar{z}_k)_I \\ (5.35) \quad &= \underbrace{((\bar{q}_k - \bar{q})g, \bar{z})_I}_{(a_1)} + \underbrace{((\bar{q} - \bar{q}_k)g, \bar{z}_k)_I}_{(a_2)}. \end{aligned}$$

We now analyze the two terms separately, ultimately highlighting their dependence from the $L^\infty(I, H_0^1(\Omega))$ -norm estimate expressed in Theorem 4.3.6.

- (a₁) In a first step, we consider the continuous state equation (2.22) with right-hand side given by $q = \bar{q}_k - \bar{q}$ and test function $\varphi = \bar{z}$. Then, observing that $(u(\bar{q}) - \bar{u})(0) = 0$, we get

$$(5.36) \quad ((\bar{q}_k - \bar{q})g, \bar{z})_I = (\partial_t(u(\bar{q}_k) - \bar{u}), \bar{z})_I + (\nabla(u(\bar{q}_k) - \bar{u}), \nabla \bar{z})_I.$$

The idea is to choose conveniently the test function in the adjoint equation (5.3b) in order to have its left-hand side coinciding with the right-hand side of (5.36). This is accomplished with the choice $\varphi = u(\bar{q}_k) - \bar{u}$, which leads us to

$$(5.37) \quad ((\bar{q}_k - \bar{q})g, \bar{z})_I = (\bar{u} - u_d, u(\bar{q}_k) - \bar{u})_I + 2\langle (\nabla \bar{u} \nabla(u(\bar{q}_k) - \bar{u}), \omega), \bar{\mu} \rangle.$$

We focus our attention on the second term in the right-hand side of (5.37), where we will exploit the complementary slackness condition (5.3d) and the positivity of the Lagrange multiplier $\bar{\mu}$. Indeed, performing firstly some easy algebraic manipulation, we have

$$\begin{aligned} &2\langle (\nabla \bar{u} \nabla(u(\bar{q}_k) - \bar{u}), \omega), \bar{\mu} \rangle \\ &= \langle (|\nabla \bar{u}|^2 + |\nabla u(\bar{q}_k)|^2, \omega), \bar{\mu} \rangle - 2\langle (|\nabla \bar{u}|^2, \omega), \bar{\mu} \rangle \\ (5.38) \quad &= \langle (|\nabla u(\bar{q}_k)|^2 - |\nabla \bar{u}|^2, \omega), \bar{\mu} \rangle \\ &= \langle (|\nabla u(\bar{q}_k)|^2 - |\nabla \bar{u}_k|^2 + |\nabla \bar{u}_k|^2 - |\nabla \bar{u}|^2, \omega), \bar{\mu} \rangle \\ &\leq \langle (|\nabla u(\bar{q}_k)|^2 - |\nabla \bar{u}_k|^2, \omega), \bar{\mu} \rangle + \langle b - F(\bar{u}), \bar{\mu} \rangle, \end{aligned}$$

using in the last step the feasibility of \bar{u}_k . As anticipated, we observe that the last term in (5.38) displays the complementary slackness condition (5.3d). Then, the Cauchy-Schwarz inequality and the boundedness in $L^\infty(I, H_0^1(\Omega))$ of $u(\bar{q}_k)$ and \bar{u}_k , yield

$$\begin{aligned} &\langle (|\nabla u(\bar{q}_k)|^2 - |\nabla \bar{u}_k|^2, \omega), \bar{\mu} \rangle + \langle b - F(\bar{u}), \bar{\mu} \rangle \\ &\leq \|\bar{\mu}\|_{C(\bar{I})^*} \|\omega\|_{L^\infty(\Omega)} \| |\nabla u(\bar{q}_k)|^2 - |\nabla \bar{u}_k|^2 \|_{L^\infty(I, L^2(\Omega))} \\ (5.39) \quad &\leq c \| (|\nabla u(\bar{q}_k)| - |\nabla \bar{u}_k|) (|\nabla u(\bar{q}_k)| + |\nabla \bar{u}_k|) \|_{L^\infty(I, L^2(\Omega))} \\ &\leq c \| |\nabla u(\bar{q}_k)| - |\nabla \bar{u}_k| \|_{L^\infty(I, L^2(\Omega))} \\ &\leq c \| \nabla(u(\bar{q}_k) - \bar{u}_k) \|_{L^\infty(I, L^2(\Omega))} \\ &= c \| u(\bar{q}_k) - \bar{u}_k \|_{L^\infty(I, H_0^1(\Omega))} \end{aligned}$$

Therefore, combining (5.38) with (5.39) in (5.37) we obtain the following estimate for (a_1)

$$(5.40) \quad ((\bar{q}_k - \bar{q})g, \bar{z})_I \leq (\bar{u} - u_d, u(\bar{q}_k) - \bar{u})_I + c\|u(\bar{q}_k) - \bar{u}_k\|_{L^\infty(I, H_0^1(\Omega))}.$$

(a_2) We proceed along the same lines of the previous case using the semi-discrete and adjoint equation in place of the continuous one. In particular, we consider (2.23) with right-hand side $q = \bar{q} - \bar{q}_k$ and $\varphi = \bar{z}_k$. Then, observing that $(u_k(\bar{q}) - \bar{u}_k)(0) = 0$, through the choice $\varphi = u_k(\bar{q}) - \bar{u}_k$ in (5.7b), we have

$$\begin{aligned} ((\bar{q} - \bar{q}_k)g, \bar{z}_k)_I &= B(u_k(\bar{q}) - \bar{u}_k, \bar{z}_k) \\ &= (\bar{u}_k - u_d, u_k(\bar{q}) - \bar{u}_k)_I + 2\langle (\nabla \bar{u}_k \nabla (u_k(\bar{q}) - \bar{u}_k), \omega), \bar{\mu}_k \rangle. \end{aligned}$$

For the second term in the right-hand side we proceed as in (5.38). Thanks to the semi-discrete complementary condition (5.7d), the feasibility of \bar{u} and the boundedness of $\|\bar{\mu}_k\|_{C(\bar{I})^*}$ stated in Lemma 5.2.2, we obtain

$$\begin{aligned} &2\langle (\nabla \bar{u}_k \nabla (u_k(\bar{q}) - \bar{u}_k), \omega), \bar{\mu}_k \rangle \\ &= \langle (|\nabla \bar{u}_k|^2 + |\nabla u_k(\bar{q})|^2, \omega), \bar{\mu}_k \rangle - 2\langle (|\nabla \bar{u}_k|^2, \omega), \bar{\mu}_k \rangle \\ &\leq \langle (|\nabla u_k(\bar{q})|^2 - |\nabla \bar{u}|^2, \omega), \bar{\mu}_k \rangle + \langle b - F(\bar{u}_k), \bar{\mu}_k \rangle \\ &\leq \|\bar{\mu}_k\|_{C(\bar{I})^*} \|\omega\|_{L^\infty(\Omega)} \| |\nabla u_k(\bar{q})|^2 - |\nabla \bar{u}|^2 \|_{L^\infty(I, L^2(\Omega))} \\ &\leq c\|u_k(\bar{q}) - \bar{u}\|_{L^\infty(I, H_0^1(\Omega))} \end{aligned}$$

Thus, we conclude (a_2) with the estimate

$$(5.41) \quad ((\bar{q} - \bar{q}_k)g, \bar{z}_k)_I \leq (\bar{u}_k - u_d, u_k(\bar{q}) - \bar{u}_k)_I + c\|u_k(\bar{q}) - \bar{u}\|_{L^\infty(I, V)}.$$

Going back to (5.35) and combining (5.40) with (5.41), we have

$$(5.42) \quad \alpha\|\bar{q} - \bar{q}_k\|_{L^2(I), \mathbb{R}^m}^2 \leq (\bar{u} - u_d, u(\bar{q}_k) - \bar{u})_I + (\bar{u}_k - u_d, u_k(\bar{q}) - \bar{u}_k)_I + c(\|u(\bar{q}_k) - \bar{u}_k\|_{L^\infty(I, V)}\|u_k(\bar{q}) - \bar{u}\|_{L^\infty(I, V)}).$$

We note that

$$\|\bar{u} - \bar{u}_k\|_I^2 = (\bar{u} - u_d, \bar{u} - \bar{u}_k) - (\bar{u}_k - u_d, \bar{u} - \bar{u}_k),$$

which, summed to (5.42), yields

$$\begin{aligned} \|\bar{u} - \bar{u}_k\|_I^2 + \alpha\|\bar{q} - \bar{q}_k\|_{L^2(I, \mathbb{R}^m)}^2 &\leq (\bar{u} - u_d, u(\bar{q}_k) - \bar{u}_k)_I + (\bar{u}_k - u_d, u_k(\bar{q}) - \bar{u})_I \\ &\quad + c(\|u(\bar{q}_k) - \bar{u}_k\|_{L^\infty(I, H_0^1(\Omega))}\|u_k(\bar{q}) - \bar{u}\|_{L^\infty(I, H_0^1(\Omega))}) \\ &\leq c\left(\|u(\bar{q}_k) - \bar{u}_k\|_I + \|u_k(\bar{q}) - \bar{u}\|_I \right. \\ &\quad \left. + \|u(\bar{q}_k) - \bar{u}_k\|_{L^\infty(I, H_0^1(\Omega))}\|u_k(\bar{q}) - \bar{u}\|_{L^\infty(I, H_0^1(\Omega))}\right). \end{aligned}$$

Then, in the relation above we bound the $\|\cdot\|_I$ -norm with the $\|\cdot\|_{L^\infty(I, H_0^1(\Omega))}$ -norm, to conclude with the help of Theorem 4.3.6

$$(5.43) \quad \|\bar{u} - \bar{u}_k\|_I^2 + \alpha\|\bar{q} - \bar{q}_k\|_{L^2(I)}^2 \leq Ck\left(\log \frac{T}{k} + 1\right)^{\frac{1}{2}},$$

which in turn gives the claim. \square

Proceeding as before, we firstly show that the discrete optimal pair and the corresponding Lagrange multiplier are bounded.

Lemma 5.2.4. *Let $(\bar{q}_{kh}, \bar{u}_{kh}) \in Q_{ad} \times U_{kh}$ be the optimal pair for Problem 2.25 with associated Lagrange multiplier $\bar{\mu}_{kh} \in C(\bar{I})^*$. Then, for k and h sufficiently small there holds*

$$(5.44) \quad \|\bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)} + \|\bar{u}_{kh}\|_I + \|\bar{\mu}_{kh}\|_{C(\bar{I})^*} \leq C.$$

Proof. The proof employed same argument as [61, Lemma 6.4] with the only difference given by the use of the $L^\infty(I, H_0^1(\Omega))$ -norm estimate of Theorem 4.3.13 instead of the $L^\infty(I, L^2(\Omega))$ -norm. \square

After this preparation, we analyze the convergence property between the semi-discrete and discrete problem, concluding the proof of Theorem 5.2.1. This result has been presented in [56, Theorem 5.3]

Theorem 5.2.5. *Let $\bar{q}_k \in Q_{ad}$ be the optimal control of the semi-discrete problem (2.25) and $\bar{q}_{kh} \in Q_{ad}$ be the one of the discrete problem (2.26). Then there holds the following error estimate*

$$(5.45) \quad \alpha \|\bar{q}_k - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)}^2 \leq Ch.$$

Proof. The proof moves along the same lines of the error estimate for the semi-discrete case in Theorem 5.2.3. In particular, we test the semi-discrete and discrete variational inequality (5.7c) and (5.9c) with $q = \bar{q}_{kh}$ and $q = \bar{q}_k$, respectively. Then, adding the resulting inequalities we have

$$(5.46) \quad \begin{aligned} \alpha \|\bar{q}_k - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)} &\leq (\bar{z}_k - \bar{z}_{kh}, (\bar{q}_{kh} - \bar{q}_k)g)_I \\ &= \underbrace{(\bar{z}_k, (\bar{q}_{kh} - \bar{q}_k)g)_I}_{(a_1)} + \underbrace{(\bar{z}_{kh}, (\bar{q}_k - \bar{q}_{kh})g)_I}_{(a_2)}. \end{aligned}$$

and we analyze the two terms separately.

- (a₁) The idea is to express (a₁) in term of the semi-discrete state equation (2.23) and the semi-discrete adjoint equation (5.7b), equalizing the term containing the bilinear form $B(\cdot, \cdot)$.

First, we consider (2.23) with right-hand side $q_k = \bar{q}_{kh} - \bar{q}_k$ and we set $\varphi = \bar{z}_k$. Then, with the choice $\varphi = u_k(\bar{q}_{kh}) - \bar{u}_k$ in (5.7b), we have

$$((\bar{q}_{kh} - \bar{q}_k)g, \bar{z}_k)_I = (\bar{u}_k - u_d, u_k(\bar{q}_{kh}) - \bar{u}_k)_I + \langle 2(\nabla \bar{u}_k \nabla(u_k(\bar{q}_{kh}) - \bar{u}_k), \omega), \bar{\mu}_k \rangle.$$

As in Theorem 5.2.3, in view of (5.7d) we rewrite conveniently the term containing the Lagrange multiplier $\bar{\mu}_k$. The feasibility of \bar{u}_{kh} and $\bar{\mu}_k$ being bounded, lead to

$$(5.47) \quad \begin{aligned} &2\langle (\nabla \bar{u}_k \nabla(u_k(\bar{q}_{kh}) - \bar{u}_k), \omega), \bar{\mu}_k \rangle \\ &= \langle (|\nabla \bar{u}_k|^2 + |\nabla u_k(\bar{q}_{kh})|^2, \omega), \bar{\mu}_k \rangle - 2\langle (|\nabla \bar{u}_k|^2, \omega), \bar{\mu}_k \rangle \\ &= \langle (|\nabla u_k(\bar{q}_{kh})|^2 - |\nabla \bar{u}_k|^2, \omega), \bar{\mu}_k \rangle \\ &= \langle (|\nabla u_k(\bar{q}_{kh})|^2 - |\nabla \bar{u}_{kh}|^2 + |\nabla \bar{u}_{kh}|^2 - |\nabla \bar{u}_k|^2, \omega), \bar{\mu}_k \rangle \\ &\leq \langle (|\nabla u_k(\bar{q}_{kh})|^2 - |\nabla \bar{u}_{kh}|^2, \omega), \bar{\mu}_k \rangle + \langle b - F(\bar{u}_k), \bar{\mu}_k \rangle, \\ &\leq \|\bar{\mu}_k\|_{C(I)^*} \|\omega\|_{L^\infty(\Omega)} \| |\nabla u_k(\bar{q}_{kh})|^2 - |\nabla \bar{u}_{kh}|^2 \|_{L^\infty(I, L^2(\Omega))} \\ &\leq c \|u_k(\bar{q}_{kh}) - \bar{u}_{kh}\|_{L^\infty(I, H_0^1(\Omega))} \end{aligned}$$

Therefore, for (a_1) we conclude

$$(5.48) \quad (\bar{z}_k, (\bar{q}_{kh} - \bar{q}_k))_I \leq (\bar{u}_k - u_d, u_k(\bar{q}_{kh}) - \bar{u}_k)_I + c \|u_k(\bar{q}_{kh}) - \bar{u}_{kh}\|_{L^\infty(I, H_0^1(\Omega))}.$$

(a_2) We now use the discrete state equation (5.9a) with $\varphi = \bar{z}_{kh}$ and right-hand side given by $q_{kh} = \bar{q}_k - \bar{q}_{kh}$, together with the discrete adjoint equation (5.9b) with $\varphi = u_{kh}(\bar{q}_k) - \bar{u}_{kh}$. Proceeding as before, this setting leads to (5.49)

$$\begin{aligned} (\bar{z}_{kh}, (\bar{q}_k - \bar{q}_{kh})g)_I &\leq (\bar{u}_{kh} - u_d, u_{kh}(\bar{q}_k) - \bar{u}_{kh})_I \\ &\quad + c \|\bar{\mu}_{kh}\|_{C(\bar{I})^*} \|\omega\|_{L^\infty(\Omega)} \|\bar{u}_k - u_{kh}(\bar{q}_k)\|_{L^\infty(I, H_0^1(\Omega))} \\ &\leq (\bar{u}_{kh} - u_d, u_{kh}(\bar{q}_k) - \bar{u}_{kh})_I + c \|\bar{u}_k - u_{kh}(\bar{q}_k)\|_{L^\infty(I, H_0^1(\Omega))}. \end{aligned}$$

using Lemma 5.2.4 to bound $\bar{\mu}_{kh}$.

We combine relation (5.48) with (5.49) in (5.46) and obtain

$$(5.50) \quad \begin{aligned} \alpha \|\bar{q}_k - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)} &\leq (\bar{u}_k - u_d, u_k(\bar{q}_{kh}) - \bar{u}_k)_I + (\bar{u}_{kh} - u_d, u_{kh}(\bar{q}_k) - \bar{u}_{kh})_I \\ &\quad + c (\|u_k(\bar{q}_{kh}) - \bar{u}_{kh}\|_{L^\infty(I, H_0^1(\Omega))} + \|\bar{u}_k - u_{kh}(\bar{q}_k)\|_{L^\infty(I, H_0^1(\Omega))}). \end{aligned}$$

Similar to Theorem 5.2.3, we observe that

$$\|\bar{u}_k - \bar{u}_{kh}\|_I^2 = (\bar{u}_k - u_d, \bar{u}_k - \bar{u}_{kh}) - (\bar{u}_{kh} - u_d, \bar{u}_k - \bar{u}_{kh}),$$

which, added to (5.50), yields

$$\begin{aligned} \|\bar{u}_k - \bar{u}_{kh}\|_I^2 + \alpha \|\bar{q}_k - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)}^2 &\leq \|u_k(\bar{q}_{kh}) - \bar{u}_{kh}\|_I + \|\bar{u}_k - u_{kh}(\bar{q}_k)\|_I \\ &\quad + \|u_k(\bar{q}_{kh}) - \bar{u}_{kh}\|_{L^\infty(I, H_0^1(\Omega))} + \|\bar{u}_k - u_{kh}(\bar{q}_k)\|_{L^\infty(I, H_0^1(\Omega))}. \end{aligned}$$

Then the claim follows thanks to the $L^\infty(I, H_0^1(\Omega))$ -norm estimate stated in Theorem 4.3.13 \square

5.3 Convergence analysis for non-convex optimization problems

This section is devoted to the derivation of the convergence rate of a discrete solution of Problem 2.32 with respect to a local continuous solution of Problem 2.27.

The main result, whose proof is given at the end of Section 5.3.1 with Proposition 5.3.8, is the following.

Theorem 5.3.1. *Let \bar{q} be a local solution of Problem 2.27 satisfying the assumptions of Theorem 5.1.9 and Assumption 5.1.3. Then, for k, h sufficiently small, there exists a sequence (\bar{q}_{kh}) of local solution of Problem 2.32 converging to \bar{q} as $k, h, \rightarrow 0$. Further there holds the error estimate*

$$(5.51) \quad \|\bar{q} - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)}^2 \leq c \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right).$$

This is indeed what one expects considering the same problem with a linear state equation, compare with [61, Theorem 6.1]

5.3.1 Zero-order integral constraints pointwise in time

We start the convergence analysis introducing continuous and discrete auxiliary problems in a neighborhood of a selected optimal local solution \bar{q} . Then, based on the linearized Slater point from Assumption 5.1.2, we build sequences of feasible controls (competitors) ensuring the existence of a global solution for the discrete auxiliary problems. In a final step, we show that such global solution coincides with a local solution of the original problem.

Let $r > 0$ denote a radius, to be chosen conveniently later. Recalling that $G_{kh} = F_{kh} \circ S_{kh}$, we define the sets

$$\begin{aligned} Q^r &:= \{q \in Q_{\text{ad}} \mid \|q - \bar{q}\|_{L^2(I, \mathbb{R}^m)} \leq r\}, \\ Q_{\text{feas}}^r &:= \{q \in Q^r \mid G(q) \leq 0\}, \\ Q_{kh, \text{feas}}^r &:= \{q_{kh} \in Q^r \mid G_{kh}(q) \leq 0\}, \end{aligned}$$

introduce the continuous auxiliary problem

$$(\mathbb{P}^r) \quad \min j(q) := J(q, S(q)) \quad \text{s.t. } q \in Q_{\text{feas}}^r,$$

and the discrete one

$$(\mathbb{P}_{kh}^r) \quad \min j_{kh}(q_{kh}) := J(q_{kh}, S_{kh}(q_{kh})) \quad \text{s.t. } q_{kh} \in Q_{kh, \text{feas}}^r.$$

We remark again that the control is not discretized, the index k, h is taken only to clarify the association to Problem (\mathbb{P}_{kh}^r) .

For the auxiliary problems, we assume that the Slater's point q_γ from Assumption 5.1.2 lies in the vicinity of the selected local solution.

Assumption 5.3.1. *Let $\bar{q} \in Q_{\text{feas}}$ be a selected local solution of Problem 2.27. Then for the Slater's point q_γ satisfying (5.11) it holds*

$$(5.53) \quad \|q_\gamma - \bar{q}\|_{L^2(I, \mathbb{R}^m)} \leq \frac{r}{2}.$$

Remark 5.3.2. As observed in [66, Section 2], the fact that q_γ lies in a neighborhood of \bar{q} is a reasonable assumption. This can be achieved defining

$$q_\gamma^r = \bar{q} + t(q_\gamma - \bar{q})$$

with a parameter $\gamma(r) = t\gamma \simeq r\gamma$ where

$$t = \min \left\{ 1, \frac{r}{2\|q_\gamma - \bar{q}\|} \right\},$$

which in turn gives that (5.11) holds with $t\gamma$ in place of γ .

Further, in view of the forthcoming Lemma 5.3.6, it is reasonable to assume that the same holds also for the discrete problem, namely

$$(5.54) \quad \|q_\gamma - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)} \leq \frac{r}{2}.$$

In a next step, we give an accessory result which will be repeatedly used in the rest of this section

Lemma 5.3.3. Let $B_{\frac{r}{2}}(\bar{q})$ denote an $L^2(I, \mathbb{R}^m)$ ball centered in \bar{q} with radius $\frac{r}{2}$ and let $q \in Q_{ad}$. We define three constants c_1, c_2, c_3 such that

$$\begin{aligned} \sup_{q \in B_{\frac{r}{2}}(\bar{q})} \|(u_{kh}(q) - u(q), \omega)\|_{L^\infty(I)} &\leq c_1 \left(k \left(\ln \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\ln \frac{T}{k} + 1 \right) \right), \\ \sup_{q \in B_{\frac{r}{2}}(\bar{q})} \|G''(q)\|_{\mathcal{L}(L^2(I, \mathbb{R}^m)^2; L^\infty(I))}, \sup_{q \in B_{\frac{r}{2}}(\bar{q})} \|G''_{kh}(q)\|_{\mathcal{L}(L^2(I, \mathbb{R}^m)^2; L^\infty(I))} &\leq c_2, \\ \sup_{q \in B_{\frac{r}{2}}(\bar{q})} \|(G'_{kh}(q) - G'(\bar{q}))(q_\gamma - \bar{q})\|_{L^\infty(I)} &\leq c_3 \left(k \left(\ln \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\ln \frac{T}{k} + 1 \right) \right. \\ &\quad \left. + \frac{r^2}{2} \right), \end{aligned}$$

These constants are independent of the discretization parameters k, h and remain bounded as $r \rightarrow 0$.

Proof. The independence of c_1 and c_2 are immediate. Indeed, for the former it follows from the discretization error estimates of Theorem 4.4.9 and 4.4.18. For the latter it is a consequence of the functional G being of class C^2 and a discretization bound for G'' which follows thanks to the boundedness of $\partial_u d(\cdot, \cdot, u)$ and $\partial_u^2 d(\cdot, \cdot, u)$.

For the constant c_3 , we need few more steps. Firstly, we note that

$$F(\varphi) = F_{kh}(\varphi) = \int_{\Omega} \varphi(t, x) \omega(x) dx, \quad \varphi \in W(0, T) \cup U_{kh}$$

is linear. Consequently the error satisfies

$$\begin{aligned} (G'_{kh}(q) - G'(\bar{q}))(q_\gamma - \bar{q}) &= F_{kh}(S'_{kh}(q)(q_\gamma - \bar{q})) - F(S'(\bar{q})(q_\gamma - \bar{q})) \\ &= \left(\omega, (S'_{kh}(q) - S'(\bar{q}))(q_\gamma - \bar{q}) \right) \\ &= \left(\omega, (S'_{kh}(q) - S'(q) + S'(q) - S'(\bar{q}))(q_\gamma - \bar{q}) \right) \\ &\leq C \left(\| (S'_{kh}(q) - S'(q))(q_\gamma - \bar{q}) \|_{L^\infty(I, H)} \right. \\ &\quad \left. + \| q - \bar{q} \|_{L^2(I, \mathbb{R}^m)} \| q_\gamma - \bar{q} \|_{L^2(I, \mathbb{R}^m)} \right), \end{aligned}$$

5.3. Convergence analysis for non-convex optimization problems

using in the last step the stability of S' , i.e., (3.31e). The term left is a discretization error and we can apply [61, Corollary 5.5, 5.11] to get

$$\begin{aligned} \|(S'_{kh}(q) - S'(q))(q_\gamma - \bar{q})\|_{L^\infty(I, H)} &\leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right) \\ &\quad \cdot (\|g\|_{L^\infty(\Omega)} \|q_\gamma - \bar{q}\|_{L^\infty(I, \mathbb{R}^m)}). \end{aligned}$$

We conclude noting that, by virtue of the control constraints, we have

$$\|q_\gamma - \bar{q}\|_{L^\infty(I, \mathbb{R}^m)} \leq |q_{\max} - q_{\min}|,$$

and therefore, thanks to (5.53), we conclude

$$|(G'_{kh}(q) - G'(\bar{q}))(q_\gamma - \bar{q})| \leq c_3 \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) + \frac{r^2}{2} \right).$$

Further, it is clear that c_1, c_2, c_3 remain bounded for $r \rightarrow 0$, because all others constants involved in the arguments above stay bounded on $B_{\frac{r}{2}}(\bar{q})$. \square

After this preparation, we summarize our requirement on the radius r relying on the following for the rest of the section.

Assumption 5.3.2. *Let the radius $r > 0$ be small enough such that the quadratic growth condition*

$$j(q) \geq j(\bar{q}) + \delta \|q - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2$$

holds for any $q \in Q_{feas}^r$.

Moreover, in view of $\gamma(r) \simeq r\gamma$, let r be small enough such that

$$(5.55) \quad -\gamma(\tilde{r}) + \left(c_2 + \frac{c_3}{2}\right) \tilde{r}^2 \leq -\frac{3}{4} \gamma(\tilde{r})$$

holds for all $\tilde{r} \leq r$.

In the following result, we build a sequence of feasible competitors for (\mathbb{P}_{kh}^r) based on the Slater's point q_γ . This will guarantee that the set $Q_{kh,feas}^r$ is not empty, leading to the existence of a global solution for the discrete auxiliary problem.

Proposition 5.3.4. *Let \bar{q} be a local solution of (\mathbb{P}) and q_γ be the Slater's point from Assumption 5.1.2. Let*

$$(5.56) \quad t(k, h) = \frac{c_1(k(\log(T/k) + 1)^{1/2} + h^2(\log(T/k) + 1))}{c_4 r^2 - \gamma}$$

be given with c_4 such that $0 < c_4 r^2 - \gamma < \gamma/2$. Then, the sequence of controls defined by

$$(5.57) \quad q_{t(k,h)} = \bar{q} + t(k, h)(q_\gamma - \bar{q})$$

is feasible for (\mathbb{P}_{kh}^r) , for k, h sufficiently small such that $0 < t(k, h) < 1$.

Proof. In a first step, we perform a Taylor's expansion of $G(q_{t(k,h)})$ at \bar{q} obtaining

$$G(q_{t(k,h)}) = G(\bar{q}) + G'(\bar{q})(q_{t(k,h)} - \bar{q}) + \frac{1}{2}G''(q_\zeta)(q_{t(k,h)} - \bar{q})^2,$$

with q_ζ being a convex combination of $q_{t(k,h)}$ and \bar{q} .

We used this expansion, in combination with the definition of $q_{t(k,h)}$, in the following calculations to obtain

$$\begin{aligned} G_{kh}(q_{t(k,h)}) &= G_{kh}(q_{t(k,h)}) - G(q_{t(k,h)}) + G(q_{t(k,h)}) \\ &= G_{kh}(q_{t(k,h)}) - G(q_{t(k,h)}) + G(\bar{q}) + G'(\bar{q})(q_{t(k,h)} - \bar{q}) \\ &\quad + \frac{1}{2}G''(q_\zeta)(q_{t(k,h)} - \bar{q})^2 \\ &= G_{kh}(q_{t(k,h)}) - G(q_{t(k,h)}) + G(\bar{q}) + t(k,h)G(\bar{q}) - t(k,h)G(\bar{q}) \\ &\quad + t(k,h)G'(\bar{q})(q_\gamma - \bar{q}) + \frac{1}{2}G''(q_\zeta)(q_{t(k,h)} - \bar{q})^2 \\ &= \underbrace{G_{kh}(q_{t(k,h)}) - G(q_{t(k,h)})}_{(a_1)} \\ &\quad + \underbrace{(1 - t(k,h))G(\bar{q}) + t(k,h)(G(\bar{q}) + G'(\bar{q})(q_\gamma - \bar{q}))}_{(a_2)} \\ &\quad + \underbrace{\frac{1}{2}G''(q_\zeta)(q_{t(k,h)} - \bar{q})^2}_{(a_3)}. \end{aligned}$$

We now analyze the three terms separately.

(a_1) The first term follows by definition of c_1

$$\begin{aligned} G_{kh}(q_{t(k,h)}) - G(q_{t(k,h)}) &= (\bar{u}_{kh}(q_{t(k,h)}) - u(q_{t(k,h)}), \omega(x))_I \\ &\leq c_1 \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right). \end{aligned}$$

(a_2) To handle this part, we exploit the feasibility of \bar{q} for Problem 2.27 and Slater's regularity condition from Assumption 5.1.2. Indeed, for k, h sufficiently small, such that $0 < t(k, h) < 1$, we have

$$\begin{aligned} (1 - t(k, h))G(\bar{q}) &\leq 0, \\ t(k, h)(G(\bar{q}) + G'(\bar{q})(q_\gamma - \bar{q})) &\leq -t(k, h)\gamma, \end{aligned}$$

and therefore

$$(a_2) \leq -t(k, h)\gamma.$$

(a_3) The definition of c_2 directly entails

$$G''(q_\zeta)(q_{t(k,h)} - \bar{q})^2 \leq c_2 t(k, h)^2 \|q_\gamma - \bar{q}\|_{L^2(I, \mathbb{R}^m)}^2 \leq c_2 t(k, h)^2 \frac{r^2}{4}.$$

We combine the three parts above and, using the definition of $t(k, h)$, we get

$$\begin{aligned} G_{kh}(q_{t(k,h)}) &\leq c_1 \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right) + t(k, h) (c_2 t(k, h) \frac{r^2}{4} - \gamma) \\ &= t(k, h) (c_4 r^2 - \gamma) + t(k, h) (c_2 t(k, h) \frac{r^2}{4} - \gamma) \\ &= t(k, h) (c_4 r^2 - 2\gamma + c_2 t(k, h) r^2). \end{aligned}$$

Then, once k, h are sufficiently small such that $0 < t(k, h) < 1$, we obtain from (5.55) and the definition of c_4 that

$$\begin{aligned} G_{kh}(q_{t(k,h)}) &\leq t(k, h) (c_4 r^2 - 2\gamma + c_2 r^2) \\ &\leq (c_4 - \gamma) + (c_2 r^2 - \gamma) \\ &\leq \frac{\gamma}{2} - \frac{3}{4}\gamma \\ &\leq -\frac{1}{4}\gamma < 0, \end{aligned}$$

ensuring the feasibility of $q_{t(k,h)}$. \square

Corollary 5.3.5. *For k, h sufficiently small, there exists at least one global solution $\bar{q}_{kh}^r \in Q_{kh,feas}^r$ of (\mathbb{P}_{kh}^r) .*

In a second step, we show that the linearized Slater's regularity condition from Assumption 5.1.2 holds for the discrete auxiliary problem (\mathbb{P}_{kh}^r) .

Lemma 5.3.6. *Under Assumption 5.1.2, for k, h sufficiently small it holds*

$$(5.58) \quad G_{kh}(\bar{q}_{kh}^r) + G'_{kh}(\bar{q}_{kh}^r)(q_\gamma - \bar{q}_{kh}^r) \leq -\frac{1}{2}\gamma < 0 \quad \text{on } \bar{I}.$$

Proof. In view of Assumption 5.1.2, we add and subtract $G(\bar{q})$, $G_{kh}(\bar{q})$, $G'(\bar{q})(q_\gamma - \bar{q})$ to the left-hand side of (5.58) obtaining

$$\begin{aligned} G_{kh}(\bar{q}_{kh}^r) + G'(\bar{q}_{kh}^r)(q_\gamma - \bar{q}_{kh}^r) &= G(\bar{q}) + G'(\bar{q})(q_\gamma - \bar{q}) + G_{kh}(\bar{q}_{kh}^r) \\ &\quad + G'(\bar{q}_{kh}^r)(q_\gamma - \bar{q}_{kh}^r) - G(\bar{q}) - G'(\bar{q})(q_\gamma - \bar{q}) \\ &\leq -\gamma + \underbrace{G_{kh}(\bar{q}_{kh}^r) + G'_{kh}(\bar{q}_{kh}^r)(\bar{q} - \bar{q}_{kh}^r) - G_{kh}(\bar{q})}_{(b_1)} \\ &\quad + \underbrace{G_{kh}(\bar{q}) - G(\bar{q})}_{(b_2)} + \underbrace{(G'_{kh}(\bar{q}_{kh}^r) - G'(\bar{q}))(q_\gamma - \bar{q})}_{(b_3)}. \end{aligned}$$

We analyze the three parts separately.

(b₁) We observe that the Taylor's expansion of $G_{kh}(\bar{q})$ at \bar{q}_{kh}^r reads

$$G_{kh}(\bar{q}) = G_{kh}(\bar{q}_{kh}^r) + G'_{kh}(\bar{q}_{kh}^r)(\bar{q} - \bar{q}_{kh}^r) + \frac{1}{2}G''_{kh}(q_\zeta)(\bar{q} - \bar{q}_{kh}^r)^2,$$

where q_ζ is convex combination of \bar{q} and \bar{q}_{kh}^r . This, in turn, gives

$$(b_1) = -\frac{1}{2}G''_{kh}(q_\zeta)(\bar{q} - \bar{q}_{kh}^r)^2,$$

implying that

$$(b_1) \leq c_2 \|\bar{q} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)}^2 \leq c_2 r^2$$

because G_{kh} is a C^2 -functional and \bar{q}_{kh}^r is feasible for (\mathbb{P}_{kh}^r) .

(b₂) This part easily follows from the definition of c_1

$$\begin{aligned} G_{kh}(\bar{q}) - G(\bar{q}) &= (u_{kh}(\bar{q}) - u(\bar{q}), \omega(x)) \\ &\leq c_1 \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right). \end{aligned}$$

(b₃) Also this part directly follows from the definition of the constant c_3

$$(G'_{kh}(\bar{q}_{kh}^r) - G'(\bar{q}))(q_\gamma - \bar{q}) \leq c_3 \left(k \left(\ln \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\ln \frac{T}{k} + 1 \right) + \frac{r^2}{2} \right)$$

We combine the estimates for parts (b₁) – (b₃) and, for k, h sufficiently small and thanks to (5.55), we conclude

$$\begin{aligned} G_{kh}(\bar{q}_{kh}^r) + G'(\bar{q}_{kh}^r)(q_\gamma - \bar{q}_{kh}^r) &\leq -\gamma + c_2 r^2 \\ &\quad + c_1 \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right) \\ &\quad + c_3 \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) + \frac{r^2}{2} \right) \\ &\leq -\gamma + \left(c_2 + \frac{c_3}{2} \right) r^2 + (c_1 + c_3) \cdot \\ &\quad \cdot \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right) \\ &\leq -\frac{3}{4} \gamma + (c_1 + c_3) \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \right. \\ &\quad \left. + h^2 \left(\log \frac{T}{k} + 1 \right) \right) \\ &\leq -\frac{1}{2} \gamma. \end{aligned}$$

□

Up to this point, we have built a sequence of feasible competitors for the discrete problem (\mathbb{P}_{kh}^r) which, in particular, has ensured the existence of a global solution $\bar{q}_{kh}^r \in Q_{kh, \text{feas}}^r$. Based on this global solution, we now build a sequence of controls feasible for the continuous problem (\mathbb{P}^r) .

In the following, we will use again $t(k, h)$ as defined in Proposition 5.3.4 but it will be denoted with $\tau(k, h)$ in order to avoid confusion when using it as subscript.

Proposition 5.3.7. *Let \bar{q}_{kh}^r be a global optimum for (\mathbb{P}_{kh}^r) and q_γ be the Slater's point from Assumption 5.1.2. Further, let*

$$\tau(k, h) = \frac{c_1(k(\log(T/k) + 1)^{1/2} + h^2(\log(T/k) + 1))}{c_4 r^2 - \gamma}$$

5.3. Convergence analysis for non-convex optimization problems

be given with a constant c_4 such that $0 < c_4 r^2 - \gamma < \gamma/2$.
Then, the sequence of controls defined by

$$(5.59) \quad q_{\tau(k,h)} = \bar{q}_{kh}^r + \tau(k,h)(q_\gamma - \bar{q}_{kh}^r)$$

is feasible for (\mathbb{P}^r) , for k, h sufficiently small.

Proof. The proof follows the lines of Proposition 5.3.4, therefore we highlight the main arguments only. In this case, we make a Taylor's expansion of $G_{kh}(q_{\tau(k,h)})$ at \bar{q}_{kh}^r . Denoting with q_ζ a convex combination of $q_{\tau(k,h)}$ and \bar{q}_{kh}^r , we obtain

$$\begin{aligned} G(q_{\tau(k,h)}) &= \underbrace{G(q_{\tau(k,h)}) - G_{kh}(q_{\tau(k,h)})}_{(b_1)} + \\ &\quad \underbrace{(1 - \tau(k,h)G_{kh}(\bar{q}_{kh}^r)) + \tau(k,h)(G_{kh}(\bar{q}_{kh}^r) + G'_{kh}(\bar{q}_{kh}^r)(q_\gamma - \bar{q}_{kh}^r))}_{(b_2)} \\ &\quad + \underbrace{\frac{1}{2}G''_{kh}(q_\zeta)(q_{\tau(k,h)} - \bar{q}_{kh}^r)}_{(b_3)} \\ &\leq \tau(k,h) \left(c_4 r^2 - 2\gamma + c_2 \tau(k,h) r^2 \right) \end{aligned}$$

where we used the definition of c_1 for (b_1) , the feasibility of \bar{q}_{kh}^r together with the discrete Slater condition from Lemma 5.3.6 in (b_2) , and G_{kh} being a C^2 -functional together with (5.54) for (b_3) .

Then, once k, h are sufficiently small such that $0 < \tau(k,h) < 1$, in addition to the prerequisite of Lemma 5.3.6, the claim follows as in Proposition 5.3.4 with

$$G(q_{\tau(k,h)}) \leq -\frac{1}{4}\gamma < 0.$$

□

The next step consists in showing that global solutions of (\mathbb{P}_{kh}^r) converge to the selected local solution of (\mathbb{P}) . To achieve this, we use the two-way feasibility in combination with the quadratic growth condition (5.24). It is here that we truly see the value of introducing the auxiliary problems. Indeed, we will see later that global solutions of (\mathbb{P}_{kh}^r) are local solutions of (\mathbb{P}_{kh}) . Therefore, the following result anticipates the sought convergence rate relative to Problem 2.27.

Proposition 5.3.8. *Let \bar{q} be a local solution for (2.27) satisfying the assumption of Theorem 5.1.9 and Assumption 5.1.3, and let \bar{q}_{kh}^r be a global solution of (\mathbb{P}_{kh}^r) . Further, let k, h be small enough such that Propositions 5.3.4 and 5.3.7 hold. Then it holds the error estimate*

$$(5.60) \quad \|\bar{q} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)}^2 \leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{h} + 1 \right) \right).$$

Proof. We consider $q_{t(k,h)}$ and $q_{\tau(k,h)}$ defined as in Proposition 5.3.4 and Proposition 5.3.7, respectively, and assume that k, h are small enough so that $0 < t(k,h), \tau(k,h) < 1$. In a first step, we exploit $q_{\tau(k,h)}$ to write

$$(5.61) \quad \|\bar{q} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)} \leq \|\bar{q} - q_{\tau(k,h)}\|_{L^2(I, \mathbb{R}^m)} + \|q_{\tau(k,h)} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)}.$$

By virtue of Proposition 5.3.7, $q_{\tau(k,h)}$ converges strongly in $L^2(I, \mathbb{R}^m)$ to \bar{q}_{kh}^r with order $\tau(k, h)$. Therefore, for the the second term in the inequality above we have

$$\|q_{\tau(k,h)} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)} \leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} \right).$$

For the first term in the right-hand side of (5.61), we exploit the feasibility of $q_{\tau(k,h)}$ for (\mathbb{P}^r) in the quadratic growth condition (5.24), to write

$$\begin{aligned} \delta \|\bar{q} - q_{\tau(k,h)}\|_{L^2(I, \mathbb{R}^m)}^2 &\leq j(q_{\tau(k,h)}) - j(\bar{q}) \\ &= j(q_{\tau(k,h)}) - j_{kh}(\bar{q}_{kh}^r) + j_{kh}(\bar{q}_{kh}^r) - j_{kh}(q_{\tau(k,h)}) \\ &\quad + j_{kh}(q_{\tau(k,h)}) - j(\bar{q}) \\ &\leq \underbrace{j(q_{\tau(k,h)}) - j_{kh}(\bar{q}_{kh}^r)}_{(c_1)} + \underbrace{j_{kh}(q_{\tau(k,h)}) - j(\bar{q})}_{(c_2)}, \end{aligned}$$

where in the last step we have used that $q_{\tau(k,h)} \in Q_{kh, \text{feas}}^r$ and \bar{q}_{kh}^r is a global optimum for (\mathbb{P}_{kh}^r) .

We now analyze the two terms separately.

(c₁) With simple algebraic manipulations and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} j(q_{\tau(k,h)}) - j_{kh}(\bar{q}_{kh}^r) &\leq \frac{1}{2} \|u(q_{\tau(k,h)}) + u_{kh}(\bar{q}_{kh}^r) - 2u_d\|_I \|u(q_{\tau(k,h)}) - u_{kh}(\bar{q}_{kh}^r)\|_I \\ &\quad + \frac{\alpha}{2} \|q_{\tau(k,h)} + \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)} \|q_{\tau(k,h)} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)}. \end{aligned}$$

Then, using the stability of the solution u and u_{kh} of (2.28) and (2.30), respectively, with the boundedness of Q_{ad} , we get with the help of the Cauchy-Schwarz inequality

$$\begin{aligned} j(q_{\tau(k,h)}) - j_{kh}(\bar{q}_{kh}^r) &\leq C \left(\|u(q_{\tau(k,h)}) - u(\bar{q}_{kh}^r)\|_I + \|u(\bar{q}_{kh}^r) - u_{kh}(\bar{q}_{kh}^r)\|_I \right. \\ &\quad \left. + \|q_{\tau(k,h)} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)} \right) \\ &\leq C \left(\|u(\bar{q}_{kh}^r) - u_{kh}(\bar{q}_{kh}^r)\|_I + \|q_{\tau(k,h)} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)} \right), \end{aligned}$$

using (3.31a) in the last step.

The first term is a discretization error that we estimate by [68, Theorems 3.3 and 4.2] together with the regularity of the solution of (2.28), obtaining

$$\|u(\bar{q}_{kh}^r) - u_{kh}(\bar{q}_{kh}^r)\|_I \leq C(k + h^2).$$

The second term follows directly from Proposition 5.3.7 and, summing up, we conclude

$$\begin{aligned} j(q_{\tau(k,h)}) - j_k(q_k^r) &\leq C \left(k + h^2 + k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right) \\ &\leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right). \end{aligned}$$

(c₂) For this part we proceed exactly as in (c₁) obtaining

$$\begin{aligned}
 j_{kh}(q_{t(k,h)}) - j(\bar{q}) &\leq \frac{1}{2} \|u_{kh}(q_{t(k,h)}) + u(\bar{q}) - 2u_d\|_I \|u_{kh}(q_{t(k,h)}) - u(\bar{q})\|_I \\
 &\quad + \frac{\alpha}{2} \|q_{t(k,h)} + \bar{q}\|_{L^2(I, \mathbb{R}^m)} \|q_{t(k,h)} - \bar{q}\|_{L^2(I, \mathbb{R}^m)} \\
 &\leq C \left(\|u_{kh}(q_{t(k,h)}) - u(q_{t(k,h)})\|_I + \|q_{t(k,h)} - \bar{q}\|_{L^2(I, \mathbb{R}^m)} \right) \\
 &\leq C \left(k \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}} + h^2 \left(\log \frac{T}{k} + 1 \right) \right).
 \end{aligned}$$

Combining (c₁) with (c₂) and inserting the resulting inequality in (5.61), we obtain the claim. \square

As already anticipated, for k, h small enough, global solutions of (\mathbb{P}_{kh}^r) are local solutions of (\mathbb{P}_{kh}) . This is readily seen observing that the constraint $\|\bar{q} - \bar{q}_{kh}^r\|_{L^2(I, \mathbb{R}^m)} \leq r$ is not active. In particular, this ensures the existence of a sequence \bar{q}_{kh} , of local solutions to (\mathbb{P}_{kh}) , converging to \bar{q} , which, ultimately, yields Theorem 5.3.1.

Remark 5.3.9. *In the previous section, we have derived a convergence rate for the convex problem with a clear separation between the error induced by the time and space discretization, reflecting the error estimates derived for the state equation. One might ask if the same can be achieved for the non-convex case discussed in the previous Section 5.3.1. While for the time discretization we can obtain, as expected,*

$$(5.62) \quad \|\bar{q} - \bar{q}_k\|_{L^2(I, \mathbb{R}^m)}^2 \leq Ck \left(\log \frac{T}{k} + 1 \right)^{\frac{1}{2}},$$

where (\bar{q}_k) is a suitable sequence of local solutions of (2.31), this is not possible for the error between the semi-discrete and discrete solution. To be more specific, this means that we cannot obtain

$$(5.63) \quad \|\bar{q}_k - \bar{q}_{kh}\|_{L^2(I, \mathbb{R}^m)}^2 \leq Ch^2 \left(\log \frac{T}{k} + 1 \right).$$

Indeed, once introducing the semi-discrete auxiliary problem

$$(\mathbb{P}_k^r) \quad \min j_k(q_k) := J(q_k, S_k(q_k)) \quad \text{s.t. } q_k \in Q_{k,feas}^r := \{q_k \in Q^r \mid G_k(q_k) \leq 0\},$$

in order to obtain (5.63) a quadratic growth condition needs to hold uniformly in k for the solution of (\mathbb{P}_k^r) . As a consequence, the SSCs from Assumption 5.1.3 needs to be transferred to the time discrete problem and, therefore, one has to analyze the convergence of the directions of the critical cone (5.22). To avoid this challenging problem, one would assume a stronger SSC not involving critical directions. This procedure is pursued, e.g., in [66, Section 5] for the case of a semi-linear elliptic state equation. However, given the semi-infinite nature of our problem, this would be a too strong requirement and thus we prefer not to go further in this direction.

6. Conclusions and outlook

This thesis has treated parabolic optimal control problems with restrictions on mean values of the state variable and its first derivative point-wise in time. Our main focus has been the analysis of the error resulting from a space-time discretization of the problem based on the dG(0)-cG(1) method. The error in the PDEs, linear and semi-linear, has been derived with a clear separation of the temporal and spatial influence. Relying on this, we derived rates of convergence for the approximated solution of the optimal control problems. We employed two different strategies depending upon the nature of the problem, convex or non-convex. For the former, first-order optimality conditions were used; the latter has required second-order conditions and the introduction of further associated problems for the localization.

In the following, we summarize our findings and point at possible extensions.

- The a priori error estimates for the PDEs were obtained via a duality argument after the introduction of auxiliary backward problems. In this technique, the error under consideration has been used as initial data for these auxiliary problems. For the linear PDE, and in relation with the gradient state constraint, we obtained an estimate in the $L^\infty(I, H_0^1(\Omega))$ -norm which has required the derivation of several estimates for the associated auxiliary backward problems, namely in the $L^1(I, H^{-1}(\Omega))$ and $H^{-3}(\Omega)$ -norms. The obtained order of convergence $\mathcal{O}(k + h)$ has reflected what we expected from the approximation properties of the discretization used, and it has been confirmed by a numerical simulation. Further, we claimed that the same convergence rate holds when considering a semi-linear PDE. The numerical findings have confirmed this conjecture which might be verified theoretically using the idea sketched in Section 4.4.3.
- The treatment of the semi-linear PDE has been performed using a similar strategy. In view of the point-wise in time state constraint, we obtained an $L^\infty(I, L^2(\Omega))$ -norm estimate based on $L^1(I, L^2(\Omega))$ and $H^{-2}(\Omega)$ -norms estimates for the auxiliary problems. Further, we derived estimates in the $L^2(I, L^2(\Omega))$ -norm for some auxiliary problems needed to linearize the problem. The derived order of convergence of $\mathcal{O}(k + h^2)$ agrees with our expectation and it has been confirmed by a numerical simulation. Additionally, it corresponds to the convergence rate obtained in [61] for a similar setting, but in presence of a linear PDE, confirming once again the efficiency of the method employed.
- Concerning the optimization problem, the rate of convergence has been derived combining the error estimates for the PDE with the optimality conditions. For the convex case, this has been done using the variational inequality together with the complementary slackness condition from the KKT-system, considering separately the temporal and spatial discretization.

The investigation of the non-convex case has required more effort and the formulation of SSCs, upon which a quadratic-growth condition has been

obtained. The presence of local solutions has required the introduction of localized auxiliary problems. Based on a linearized Slater's point, we built sequences of feasible controls leading to the existence of a global solution for the auxiliary problem. Then, relying on the quadratic-growth condition, together with the before-mentioned sequences of feasible controls, we derived the convergence rate of the global solution toward the selected local solution. Ultimately, this has provided the sought convergence rate because a global solution of the auxiliary problem is a local solution of the original one.

- The analysis for the non-convex problem has been performed in one step, without a separation of the temporal and spatial error. To have this separation, a quadratic-growth condition, and as a consequence SSCs, must be transferred to the time discrete level. To the best of the author knowledge, this has been done so far only for stronger SSCs, that is, when a cone of critical direction is not formulated. An investigation of the stability of weaker SSCs might be of interest, e.g., to show local uniqueness of local solutions.
- Thinking of the possible industrial applications, particularly of the glass cooling processes, a natural extension would be the consideration of gradient state constraint point-wise in time and space. Very recently, the required estimates for a linear PDE in the $L^\infty(I, W^{1,\infty}(\Omega))$ -norm have been obtained in [?]. Based on these, and using the technique depicted for the convex-case, one might obtain the convergence rate for the optimal control problem.

Bibliography

- [1] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev spaces*, vol. v. 140 of Pure and applied mathematics, Academic Press, Amsterdam [u.a.], 2nd ed ed., 2003.
- [2] J. APPELL AND P. ZABREJKO, *Nonlinear Superposition Operators*, Cambridge Studies in Applied Ecology and Resource Management, Cambridge University Press, 1990.
- [3] N. ARADA AND J.-P. RAYMOND, *Dirichlet boundary control of semilinear parabolic equations. II. Problems with pointwise state constraints*, Appl. Math. Optim., 45 (2002), pp. 145–167.
- [4] W. BANGERTH, R. HARTMANN, AND G. KANSCHAT, *deal.II – a general purpose object oriented finite element library*, ACM Trans. Math. Softw., 33 (2007), pp. 24/1–24/27.
- [5] J. F. BONNANS AND P. JAISSON, *Optimal control of a parabolic equation with time-dependent state constraints*, SIAM J. Control Optim., 48 (2010), pp. 4550–4571.
- [6] S. C. BRENNER AND R. L. SCOTT, *The Mathematical Theory of Finite Element Methods*, vol. 15 of Texts Appl. Math., Springer, New York, third ed., 2008.
- [7] E. CASAS, *Pontryagin’s principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.
- [8] E. CASAS, J. C. DE LOS REYES, AND F. TRÖLTZSCH, *Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints*, SIAM J. Optim., 19 (2008), pp. 616–643.
- [9] E. CASAS AND L. A. FERNÁNDEZ, *Corrigendum: Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state*, Appl. Math. Optim., 28 (1993), pp. 337–339.
- [10] ———, *Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state*, Appl. Math. Optim., 27 (1993), pp. 35–56.
- [11] E. CASAS AND M. MATEOS, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454 (electronic).
- [12] E. CASAS, M. MATEOS, AND J.-P. RAYMOND, *Pontryagin’s principle for the control of parabolic equations with gradient state constraints*, Nonlinear Anal., 46 (2001), pp. 933–956.

BIBLIOGRAPHY

- [13] E. CASAS, J.-P. RAYMOND, AND H. ZIDANI, *Pontryagin's principle for local solutions of control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1182–1203.
- [14] E. CASAS AND F. TRÖLTZSCH, *Error estimates for the finite-element approximation of a semilinear elliptic control problem*, Control Cybernet., 31 (2002), pp. 695–712. Well-posedness in optimization and related topics (Warsaw, 2001).
- [15] ———, *Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM J. Optim., 13 (2002), pp. 406–431.
- [16] E. CASAS AND F. TRÖLTZSCH, *Second order analysis for optimal control problems: improving results expected from abstract theory*, SIAM J. Optim., 22 (2012), pp. 261–279.
- [17] E. CASAS AND F. TRÖLTZSCH, *Second order optimality conditions and their role in pde control*, Jahresber. Dtsch. Math.-Ver., 117 (2015), pp. 3–44.
- [18] K. CHRYSAFINOS, *Convergence of discontinuous Galerkin approximations of an optimal control problem associated to semilinear parabolic PDE's*, M2AN Math. Model. Numer. Anal., 44 (2010), pp. 189–206.
- [19] D. CLEVER AND J. LANG, *Optimal control of radiative heat transfer in glass cooling with restrictions on the temperature gradient*, Optimal Control Appl. Meth., 33 (2012), pp. 157–175.
- [20] J. C. DE LOS REYES, P. MERINO, J. REHBERG, AND F. TRÖLTZSCH, *Optimality conditions for state-constrained PDE control problems with time-dependent controls*, Control Cybernet., 37 (2008), pp. 5–38.
- [21] K. DECKELNICK, A. GÜNTHER, AND M. HINZE, *Finite element approximation of elliptic control problems with constraints on the gradient*, Numer. Math., 111 (2009), pp. 335–350.
- [22] K. DECKELNICK AND M. HINZE, *Variational discretization of parabolic control problems in the presence of pointwise state constraints*, J. Comput. Math., 29 (2011), pp. 1–15.
- [23] P. DEUFLHARD, M. SEEBASS, D. STALLING, R. BECK, AND H.-C. HEGE, *Hyperthermia treatment planning in clinical cancer therapy: modelling, simulation, and visualization*, in Computational Physics, Chemistry and Biology, A. Sydow, ed., vol. 3, 1997.
- [24] E. DIBENEDETTO, *On the local behaviour of solutions of degenerate parabolic equations with measurable coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 13 (1986), pp. 487–535.
- [25] N. DINCULEANU, *Vector measures*, International series of monographs in pure and applied mathematics, v. 95, Dt. Verlag d. Wiss., 1967.
- [26] K. EPPLER AND F. TRÖLTZSCH, *Fast optimization methods in the selective cooling of steel*, in Online optimization of large scale systems, Springer, Berlin, 2001, pp. 185–204.

- [27] K. ERIKSSON AND C. JOHNSON, *Adaptive finite element methods for parabolic problems. I. A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.
- [28] ———, *Adaptive finite element methods for parabolic problems. II. Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$* , SIAM J. Numer. Anal., 32 (1995), pp. 706–740.
- [29] K. ERIKSSON, C. JOHNSON, AND V. THOMÉE, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 611–643.
- [30] L. C. EVANS, *Partial Differential Equations*, vol. 19 of Grad. Stud. Math., AMS, Providence, second ed., 2010.
- [31] R. S. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.
- [32] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Classics in Mathematics, Springer, Berlin, 2001.
- [33] H. GOLDBERG AND F. TRÖLTZSCH, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim., 31 (1993), pp. 1007–1025.
- [34] C. GOLL, T. WICK, AND W. WOLLNER, *Dopelib: Differential equations and optimization environment; a goal oriented software library for solving pdes and optimization problems with pdes*, Archive of Numerical Software, 5 (2017).
- [35] W. GONG AND M. HINZE, *Error estimates for parabolic optimal control problems with control and state constraints*, Comput. Optim. Appl., 56 (2013), pp. 131–151.
- [36] W. GONG, M. HINZE, AND Z. ZHOU, *A priori error analysis for finite element approximation of parabolic optimal control problems with pointwise control*, SIAM J. Control Optim., 52 (2014), pp. 97–119.
- [37] ———, *Finite element method and a priori error estimates for Dirichlet boundary control problems governed by parabolic PDEs*, J. Sci. Comput., 66 (2016), pp. 941–967.
- [38] J. A. GRIEPENTROG, *Maximal regularity for nonsmooth parabolic problems in Sobolev-Morrey spaces*, Adv. Differential Equations, 12 (2007), pp. 1031–1078.
- [39] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Monographs and studies in Mathematics, Pitman, Boston, 1. ed., 1985.
- [40] M. HINTERMÜLLER AND K. KUNISCH, *PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative*, SIAM J. Optim., 20 (2009), pp. 1133–1156.

- [41] M. HINZE, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.
- [42] M. HINZE AND C. MEYER, *Stability of semilinear elliptic optimal control problems with pointwise state constraints*, Comput. Optim. Appl., 52 (2012), pp. 87–114.
- [43] M. HINZE, R. PINNAU, M. ULBRICH, AND S. ULBRICH, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications, Springer Netherlands, 2010.
- [44] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum. III. Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.
- [45] K. ITO AND K. KUNISCH, *Lagrange Multiplier Approach to Variational Problems and Applications*, Advances in Design and Control, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2008.
- [46] P. JAMET, *Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain*, SIAM J. Numer. Anal., 15 (1978), pp. 912–928.
- [47] K. KRUMBIEGEL AND J. REHBERG, *Second order sufficient optimality conditions for parabolic optimal control problems with pointwise state constraints*, SIAM J. Control Optim., 51 (2013), pp. 304–331.
- [48] I. LASIECKA, *State constrained control problems for parabolic systems: regularity of optimal solutions*, Appl. Math. Optim., 6 (1980), pp. 1–29.
- [49] D. LEYKEKHMAN AND B. VEXLER, *Optimal a priori error estimates of parabolic optimal control problems with pointwise control*, SIAM J. Numer. Anal., 51 (2013), pp. 2797–2821.
- [50] ———, *A priori error estimates for three dimensional parabolic optimal control problems with pointwise control*, SIAM J. Control Optim., 54 (2016), pp. 2403–2435.
- [51] J.-L. LIONS, *Optimal control of systems governed by partial differential equations*, vol. 170 of Die Grundlehren der mathematischen Wissenschaften, Springer, Berlin, 1971.
- [52] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications, Vol. 1*, Grundlehren der mathematischen Wissenschaften, 1972.
- [53] ———, *Non-Homogeneous Boundary Value Problems and Applications, Vol. 2*, Grundlehren der mathematischen Wissenschaften, 1972.
- [54] F. LUDOVICI, I. NEITZEL, AND W. WOLLNER, *A priori error estimates for nonstationary optimal control problems with gradient state constraints*, in PAMM, vol. 15, WILEY-VCH Verlag, 2015, pp. 611–612.

-
- [55] ———, *A priori error estimates for state constrained semilinear parabolic optimal control problems*, preprint, TU Darmstadt, 2016.
- [56] F. LUDOVICI AND W. WOLLNER, *A priori error estimates for a finite element discretization of parabolic optimization problems with pointwise constraints in time on mean values of the gradient of the state*, SIAM J. Control Optim., 53 (2015), pp. 745–770.
- [57] M. LUSKIN AND R. RANNACHER, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 93–113.
- [58] U. MACKENROTH, *Optimalitätsbedingungen und Dualität bei zustandsrestringierten parabolischen Kontrollproblemen*, Math. Operationsforsch. Statist. Ser. Optim., 12 (1981), pp. 65–89.
- [59] U. MACKENROTH, *On parabolic distributed optimal control problems with restrictions on the gradient*, Appl. Math. Optim., 10 (1983), pp. 69–95.
- [60] M. MATEOS, *Problema de Control Óptimo Gobernados por Ecuaciones Semilineales con Restricciones de Tipo Integrale sobre el Gradiente del Estado*, PhD thesis, Universidad de Cantabria, 2000.
- [61] D. MEIDNER, R. RANNACHER, AND B. VEXLER, *A priori error estimates for finite element discretizations of parabolic optimization problems with pointwise state constraints in time*, SIAM J. Control Optim., 49 (2011), pp. 1961–1997.
- [62] D. MEIDNER AND B. VEXLER, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems. I. Problems without control constraints*, SIAM J. Control Optim., 47 (2008), pp. 1150–1177.
- [63] ———, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems. II. Problems with control constraints*, SIAM J. Control Optim., 47 (2008), pp. 1301–1329.
- [64] C. MEYER, *Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints*, Control Cybernet., 37 (2008), pp. 51–83.
- [65] C. MEYER AND P. PHILLIP, *Optimizing the temperature profile during sublimation growth of sic single crystals: control of heating power, frequency, and coil position*, Crystal Growth & Design, 5 (2005), pp. 1145–1156.
- [66] I. NEITZEL, J. PFEFFERER, AND A. RÖSCH, *Finite element discretization of state-constrained elliptic optimal control problems with semilinear state equation*, SIAM J. Control Optim., 53 (2015), pp. 874–904.
- [67] I. NEITZEL AND F. TRÖLTZSCH, *On regularization methods for the numerical solution of parabolic control problems with pointwise state constraints*, ESAIM Control Optim. Calc. Var., 15 (2009), pp. 426–453.

BIBLIOGRAPHY

- [68] I. NEITZEL AND B. VEXLER, *A priori error estimates for space-time finite element discretization of semilinear parabolic optimal control problems*, Numer. Math., 120 (2012), pp. 345–386.
- [69] J. NEČAS, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie, Éditeurs, Paris; Academia, Éditeurs, Prague, 1967.
- [70] R. H. NOCHETTO, *Sharp L^∞ -error estimates for semilinear elliptic problems with free boundaries*, Numer. Math., 54 (1988), pp. 243–255.
- [71] C. ORTNER AND W. WOLLNER, *A priori error estimates for optimal control problems with pointwise constraints on the gradient of the state*, Numer. Math., 118 (2011), pp. 587–600.
- [72] R. PINNAU, *Analysis of optimal boundary control for radiative heat transfer modeled by the SP_1 -system*, Commun. Math. Sci., 5 (2007), pp. 951–969.
- [73] R. PINNAU AND G. THÖMMES, *Optimal boundary control of glass cooling processes*, Math. Methods Appl. Sci., 27 (2004), pp. 1261–1281.
- [74] R. RANNACHER, *L^∞ -stability estimates and asymptotic error expansion for parabolic finite element equations*, in Extrapolation and defect correction (1990), vol. 228 of Bonner Math. Schriften, Univ. Bonn, Bonn, 1991, pp. 74–94.
- [75] J. P. RAYMOND, *Nonlinear boundary control of semilinear parabolic problems with pointwise state constraints*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 341–370.
- [76] J.-P. RAYMOND AND F. TRÖLTZSCH, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 431–450.
- [77] J. P. RAYMOND AND H. ZIDANI, *Pontryagin’s principle for state-constrained control problems governed by parabolic equations with unbounded controls*, SIAM J. Control Optim., 36 (1998), pp. 1853–1879.
- [78] A. RÖSCH, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw., 21 (2006), pp. 121–134.
- [79] S. SALSA, *Partial differential equations in action : from modelling to theory*, Universitext, Springer, Milan, 2008.
- [80] D. SCHÖTZAU, *hp-DGFEM for Parabolic Evolution Problems. Applications to diffusion and viscous incompressible fluid flow.*, PhD thesis, Swiss Federal Institute of Technology Zürich, 1999.
- [81] A. SPRINGER AND B. VEXLER, *Third order convergent time discretization for parabolic optimal control problems with control constraints*, Comput. Optim. Appl., 57 (2014), pp. 205–240.
- [82] G. STAMPACCHIA, *Équations elliptiques du second ordre á coefficients discontinus*, Séminaire Jean Leray, (1963-1964), pp. 1–77.

- [83] ———, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.
- [84] V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, vol. 25 of Springer Ser. Comput. Math., Springer, Berlin, second ed., 2006.
- [85] F. TRÖLTZSCH, *Optimal Control of Partial Differential Equations. Theory, Methods and Applications.*, vol. 112 of Grad. Stud. Math., AMS, 2010.
- [86] F. TRÖLTZSCH, R. LEZIUS, R. KRENGEL, AND H. WEHAGE, *Mathematische Behandlung der optimalen Steuerung von Abkühlungsprozessen bei Profilstählen*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1997, pp. 513–522.
- [87] A. UNGER AND F. TRÖLTZSCH, *Fast solution of optimal control problems in the selective cooling of steel*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 447–456.
- [88] G. WANG AND X. YU, *Error estimates for an optimal control problem governed by the heat equation with state and control constraints*, Int. J. Numer. Anal. Model., 7 (2010), pp. 30–65.
- [89] M. F. WHEELER, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.
- [90] W. WOLLNER, *A priori error estimates for optimal control problems with constraints on the gradient of the state on nonsmooth polygonal domains*, in Control and Optimization with PDE Constraints, vol. 164 of Internat. Ser. Numer. Math., Springer Basel, 2013, pp. 193–215.
- [91] J. ZOWE AND S. KURCYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

CV

- Oct. 2015 - Aug. 2017
PhD candidate at TU Darmstadt, Department of Nonlinear Optimization.
- Oct. 2012 - Sept. 2015
PhD candidate at Uni Hamburg, Department of Optimization and Approximation
- Sept. 2010 - Sept. 2012
Master in Mathematical Engineering at University of Nice Sophia Antipolis and University of L'Aquila.
Part of the MSc Erasmus Mundus programme Mathmods.
- Sept. 2008 - May 2009
Bachelor (Hons) in Commercial Mathematics and Statistics at York University
Joint International dual degree programme.
- Oct. 2005 - July 2010
Bachelor in Mathematics at University of L'Aquila